UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

TRABAJO MONOGRÁFICO DE ACTUALIZACIÓN

HERRAMIENTAS BIOINFORMÁTICAS PARA LA IDENTIFICACIÓN Y ANÁLISIS DE VIRUS

QUE PARA OBTENER EL TÍTULO DE

QUÍMICO FARMACÉUTICO BIÓLOGO

PRESENTA

JESÚS ALBERTO NEQUIS GONZÁLEZ

CDMX AÑO 2023

JURADO ASIGNADO:

PRESIDENTE: Profesor: DIMITROVA DINKOVA TZVETANKA

VOCAL: Profesor: TIRADO MENDOZA GABRIELA

SECRETARIO: Profesor: HUGO GILDARDO CASTELÁN SÁNCHEZ

1er. SUPLENTE: Profesor: MEDINA FRANCO JOSE LUIS

2° SUPLENTE: Profesor: TORRES FLORES ALEJANDRO

SITIO DONDE SE DESARROLLÓ EL TEMA:
Grupo de Genómica y Dinámica Evolutiva de Microorganismos Emergentes; Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT).

ASESOR DEL TEMA:
HUGO GILDARDO CASTELÁN SÁNCHEZ

SUPERVISOR TÉCNICO:
GAMALIEL LÓPEZ LEAL

SUSTENTANTE:
JESÚS ALBERTO NEQUIS GONZÁLEZ

AGRADECIMIENTOS

A mi hijo, Marcelo Nicolai Nequis, que en todo momento estuvo, está y estará conmigo, él siempre será mi motivo para superarme día a día

A mis papás, por apoyarme en toda mi carrera universitaria y forjarme como todo un profesionista

“Porque al hombre que le agrada, Dios le da sabiduría, ciencia y gozo; más al pecador da el trabajo de recoger y amontonar, para darlo al que agrada a Dios. También esto es vanidad y aflicción de espíritu.”

Eclesiastés 2:26

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Los virus son entidades biológicas abundantes en la biosfera, capaces de infectar una amplia variedad de grupos poblacionales. Estos agentes microscópicos tienen la capacidad desencadenar una diversidad de enfermedades, debido a que su patogenicidad incluye cuadros entéricos, o enfermedades exantemáticas como lo son la varicela o sarampión, también pueden causar enfermedades localizadas y sistemáticas que van acompañadas de síntomas leves como lo son fiebre o escalofríos hasta síntomas graves, además de iniciar brotes locales, epidemias e incluso pandemias. Una de las características distintivas de los virus es su alta tasa de mutación, lo que, combinado con el crecimiento exponencial de la población humana, aumenta el riesgo de zoonosis y provoca brotes de manera más frecuente.
En la actualidad, la necesidad de estudiar a los virus desde una perspectiva genómica se hace cada vez más apremiante. Este trabajo tiene como objetivo presentar y conceptualizar diversas bases de datos y herramientas bioinformáticas para el análisis y caracterización de los virus de eucariontes y procariontes. El enfoque genómico permite obtener una visión más completa de la evolución de los virus y su interacción con hospederos.

La información genómica de los virus es valiosa para entender cómo surgen nuevas variantes, cómo se propagan y cómo interactúan con diferentes huéspedes. La bioinformática, como disciplina que combina la biología y la informática, se convierte en una aliada indispensable para procesar, analizar y extraer conocimiento de los grandes volúmenes de datos generados a partir de secuencias genómicas virales.

El acceso a bases de datos confiables y actualizadas y el acceso a herramientas bioinformáticas avanzadas permitirán a los investigadores y profesionales de la salud prepararse mejor para los desafíos que plantean los nuevos virus. Comprender las poblaciones virales y su dinámica es fundamental para desarrollar estrategias efectivas de prevención y control para proteger la salud pública en un mundo donde las enfermedades infecciosas siguen siendo una amenaza constante.

OBJETIVOS

- Revisar información disponible acerca de las diversas bases de datos y herramientas bioinformáticas en artículos científicos publicados para el análisis de virus.
- Conceptualizar técnicas tradicionales para la identificación de virus
- Visualizar y comprender diversas bases de datos para la identificación de virus actuales y emergentes
- Contribuir a la comprensión de herramientas bioinformáticas con el uso de metagenomas y establecer su importancia en la sociedad actual

PLANTEAMIENTO DEL PROBLEMA

La identificación de virus, a partir de métodos tradicionales y metagenómicos sigue siento un reto, y surgen varios desafíos y problemáticas que requieren ser abordados mediante del desarrollo de nuevas herramientas bioinformáticas y con herramientas disponibles.

Las técnicas de laboratorio clásicas para identificar y caracterizar virus presentan limitaciones, como la necesidad de cultivo previo del virus, la especificidad y sensibilidad limitadas, y el tiempo requerido para obtener resultados. Estas limitaciones pueden retrasar la detección temprana de nuevos brotes y afectar la toma de decisiones en salud pública.

Además, la información genética y epidemiológica de los virus es vasta y compleja, lo que dificulta su análisis e interpretación sin el uso de herramientas bioinformáticas adecuadas. La cantidad de secuencias genómicas disponibles y la necesidad de integrar múltiples fuentes de datos requieren soluciones bioinformáticas eficientes.

Existen diversas bases de datos que contienen información relevante sobre virus, pero acceder y gestionar adecuadamente esta información puede ser complicado para los investigadores y profesionales de la salud. La falta de una estructura integrada y de fácil acceso puede dificultar la obtención de datos relevantes.

El problema central radica en la necesidad de contar con soluciones bioinformáticas efectivas y actualizadas que permitan identificar, analizar y comprender rápidamente la diversidad de virus de eucariontes y procariontes emergentes y actuales, gestionar la información de bases de datos de manera eficiente y contribuir a una mejor toma de decisiones para enfrentar brotes epidémicos en la sociedad actual.

Capítulo 1. Definición de bioinformática y su aplicación en la identificación de virus

La bioinformática es una disciplina científica que desarrolla programas, bases de datos, algoritmos y métodos computacionales, que son incorporados en sistemas, flujos de trabajo y diversas estrategias de investigación con el objeto de estudiar y comprender los sistemas biológicos[1], desempeña un papel fundamental en la identificación de los virus, permitiendo una comprensión más profunda de su diversidad, evolución y patogénesis. Además, facilita el desarrollo de estrategias de diagnóstico, prevención y control de enfermedades virales[1]. Actualmente, se estima que la virosfera tiene una cantidad de 10^{31} de virus y con una mutación continua gracias a su diversidad genómica, por lo que la forma más prometedora, eficaz y rápida de enferntarlos es con el uso de herramientas computacionales como la bioinformática para identificar secuencias virales y sus elementos funcionales codificados para predecir y anotar sus comparar sus funciones.[2]

1.2. Importancia de la identificación de virus en la investigación médica y la salud pública

A la hora de controlar los brotes de virus es importante conocer y aprender de sus orígenes, siendo un primer paso aislarlos para su caracterización. Una problemática a nivel mundial son los virus emergentes y reemergentes debido a los brotes epidémicos que estos causan, por lo que una caracterización temprana es de vital importancia para su control. Los virus tienen una rápida evolución y son capaces de ocasionar pandemias, como el SARS-CoV-2, por lo tanto, es de suma importancia explorar y conocer la diversidad de los virus, puesto que se estima que 1.67 millones de virus aún desconocidos circulan en animales con una capacidad de transmisión zoonótica a los humanos.[3]

Dado esto, la epidemiología es un tema sustancial dentro de este ámbito, ya que la epidemiología es el estudio de la distribución, dinámica y determinantes de brotes pandémicos, ya que uno de los papeles de la epidemiología es conocer la situación de salud en diferentes grupos de población, sus determinantes y sus tendencias, identificando los problemas de salud prioritarios en la población, realizando la

vigilancia epidemiológica de enfermedades y otros problemas determinantes de la situación de salud.[4]

La vigilancia genómica proporciona una compresión más clara de los patógenos, su evolución y circulación con ayuda de datos clínicos, epidemiológicos y otras fuentes variadas, en brotes emergentes de virus con potencial para crear pandemias. Es importante aclarar que la vigilancia genómica es fundamental para poder monitorear posibles epidemias emergentes, como el caso más reciente por COVID-19, ya que se hizo un seguimiento de la propagación de las variantes y se monitorearon los cambios en el código genético de las variantes del SARS-CoV-2. En forma conjunta, esta información se utiliza para comprender mejor cómo las variantes pueden afectar a la salud pública. Por lo que la secuenciación genómica se utiliza cada vez más en la investigación ya que es una tecnología que permite conocer y descifrar el código genético que tienen todos los virus, es importante enfocarnos en ese código ya que contiene información imprescindible para su desarrollo y funcionamiento. En los últimos años, ha cobrado fuerza en incorporar capacidades de secuenciación genómica a nivel de país e integrarse con otros sistemas de vigilancia de enfermedades para garantizar que la vigilancia genómica pueda convertirse en parte de la programación nacional de salud pública.[5]

Capítulo 2. Técnicas de laboratorio tradicionales para la identificación de virus y generalidades de virus

Para tener un control efectivo sobre las pandemias futuras, es fundamental emplear técnicas de laboratorio que permitan un diagnóstico preciso. Esto se debe a que los virus conocidos transmitidos por zoonosis generalmente pueden ser estudiados utilizando cultivos celulares y/o modelos animales de experimentación.[6] Por lo que a continuación se describen los métodos tradiciones para investigar a los virus.

2.1.1. Cultivo celular

Un cultivo celular es un modelo *in vitro* de las células que pueden crecer y mantenerse en suspensión o monocapa en condiciones especiales, en un medio sintético y óptimo, simulando las condiciones del crecimiento *in vivo*, se conocen dos tipos de cultivo celular:[7]

- Cultivo primario

Es un cultivo inicial de células que son tomados directamente de los organismos recién sacrificados con un cariotipo idéntico al original, suele tener una vida útil de 7 días en promedio

- Línea celular

Provienen de un cultivo primario que fue sometido a procesos anteriormente donde se le confieren una capacidad ilimitada de multiplicación o poseen origen tumoral, suelen tener presencia de uno o más cromosomas supernumerarios, o ausencia de cromosomas que lleva a desequilibrio en la dotación cromosómica (aneuploides), proliferan de forma ilimitada.[7]

Estos cultivos son un estándar de oro para el aislamiento de virus, se desarrollan a partir de tejido y posteriormente se disgregan con 2 métodos principalmente para poder extraer las células adecuadas para el aislamiento del virus, observándose un cambio celular en la monocapa:

- Métodos mecánicos: Consiste en desprender las células del tejido por medios físicos con ayuda de instrumentos como pinzas, morteros, bisturí, entre otros, por el paso forzado del tejido a través de mallas
- Enzimáticos: Este método se recomienda para disminuir el daño que ocurre cuando se utiliza disgregación mecánica

Los cambios en las células de la monocapa indican la presencia de virus. Estos cambios en el cultivo celular se definen como el efecto citopático (ECP), que se debe a la presencia del virus, en la mayoría de los casos, el ECP aparece después de 5 a 10 días de incubación.[8]

Recordemos que el aislamiento viral es un método auxiliar e indicativo para la presencia presuntiva viral. Durante muchos años, han proporcionado un entorno deseable para la detección e identificación de muchos patógenos virales humanos. Los métodos moleculares y otros como la detección de antígenos virales, no requieren el largo período de incubación necesario para el aislamiento viral en cultivos celulares, pueden requerir menos experiencia técnica y son útiles para virus que no proliferan

en cultivos celulares estándar. La principal ventaja del enfoque de cultivo celular tradicional es la capacidad de aislar una amplia variedad de virus.[9]

Últimamente los modelos de cultivo tridimensionales (3D), han sido notablemente más sobresalientes ya que son una mejora sobre los cultivos en 2D en monocapa, en estos modelos, los cultivos celulares son inadecuadas representaciones de un microambiente, los cuales a menudo los hace predictores poco confiables *in vivo* de la eficacia y toxicidad de fármacos *in vivo*.

Los modelos 3D o esferoides 3D se asemejan más al tejido *in vivo* en términos de comunicación celular y el desarrollo de matrices extracelulares. Estas matrices ayudan a las células para sean capaces de moverse dentro de un esferoide similar a la manera en que la célula se movería en un tejido vivo.

Los esferoides son así modelos mejorados para la migración celular, diferenciación celular, supervivencia y crecimiento diferenciación, además, los cultivos celulares en 3D proporcionan una representación más exacta de la polarización celular 3D ya que en 2D, las células solo pueden ser polarizadas parcialmente, de igual manera, se evita usar los modelos animales, ya que estos modelos son relativamente más costosos y su uso ha tenido cuestiones éticas, ya que usualmente se usan estos modelos animales para proporcionar un sistema de modelo de células *in vivo* que contribuye a la comprensión tanto de la interacción virus-huésped como de los mecanismos fundamentales de los virus humanos como promover el desarrollo de medicamentos antivirales.

Las células cultivadas en una matriz 3D pueden reproducir adecuadamente la función de los tejidos 3D e imitar las interacciones célula-célula y célula-matriz *in vivo*. Los sistemas de cultivo 3D se han diseñado para permitir la investigación de patógenos infecciosos, como virus humanos, bacterias y parásitos.

En la tabla 1 se comparan los modelos 2D y 3D en cuanto a función y características celulares. Estos modelos tienen un papel invaluable en la virología actual.[10]

Tabla 1. Las diferencias de función biológica y características celulares en sistemas 2D y 3D

Características/función celular	Modelo 2D	Modelo 3D
Forma de la celda	Una sola capa	Múltiples capas
Morfología De estructuras agregadas/esferoidales	Células planas y estiradas en forma de lámina en monocapa	De estructuras agregadas/esferoidales
Polaridad	Polarización parcial	Representación más precisa de la polarización celular
Rigidez	Rigidez alta	Rigidez baja
Migración	Un solo mecanismo	Diversas estrategias de migración celular
Adherencias	Representar etapas exageradas de dinámica in vivo	Genera adherencias comparables con la adherencia 3D in vivo
Proliferación	Células tumorales crecidas en monocapa más rápido que en esferoides 3D	Similar a la situación in vivo
Expresión génica/expresión de proteínas	A menudo muestran niveles diferenciales de genes/proteínas en comparación con los modelos in vivo	Expresión de genes y proteínas in vivo para estar presente en modelos 3D

He, B.; Chen, G.; Zeng, Y. Three-Dimensional Cell Culture Models for Investigating Human Viruses. Virol. Sin. 2016, 31 (5), 363–379. https://doi.org/10.1007/s12250-016-3889-z.

Para una oportuna respuesta frente epidemias y pandemias futuras es necesario la detección pronta de agentes aislados de muestras clínicas y de pacientes de estudios, hoy en día se conocen los métodos de detección asociados con Reacción en Cadena de la Polimerasa (PCR), tecnologías CRISPR/Cas (en inglés: Clustered Regularly Interspaced Short Palindromic Repeats, en español: repeticiones palindrómicas cortas agrupadas y regularmente interespaciadas, la secuenciación de siguiente generación (NGS), inmunoensayos y ensayos basados en células, sin embargo algunos enfoques permiten que solo haya detección de ácidos nucleicos o proteínas virales, o la detección de partículas virales viables.

Es por ello que el método SHERLOK (Desbloqueo de reportero enzimático especifico de alta sensibilidad) de desarrollo, que aumenta la efectividad y cantidad de genomas virales en relación con los genomas del huésped, esta herramienta de detección de ácidos nucleicos con alta sensibilidad implica la amplificación del material del genoma viral y la escisión indirecta de las sondas de ARN marcadas con fluorescencia por la proteína Cas13, lo que permite la cuantificación del material viral en la muestra. La sensibilidad de este método resulta del hecho de que incluso cantidades insignificantes de ácidos nucleicos virales del material biológico se someten a amplificación de polimerasa de recombinasa (RPA), con un paso adicional de transcripción inversa para la detección de virus de ARN. Las copias de ADN

amplificadas de los virus se vuelven a convertir en ARN utilizando la ARN polimerasa T7 dependiente de ADN. Posteriormente, utilizando el ARN genómico gARN, la proteína Cas13 reconoce las secuencias de nucleótidos virales, lo que conduce al corte colateral de los nucleótidos marcados, lo que permite la determinación cuantitativa del contenido viral en las muestras.

Las propiedades de este reportero deben ser muy específicas del virus para permitir que los investigadores detecten específicamente el virus de interés. Existen diferentes principios para la detección de diferentes virus dependiendo de su estructura genómica y ciclo de replicación. La transactivación de estructuras virales en los vectores de algunos virus de ARN puede ocurrir a través de la infección de proteínas virales.[11]

Tabla 2. Ventajas y desventajas de los diversos métodos de detección y aplicación.

Objetivo de detección y aplicación.		Métodos	Ventajas	Desventajas
Detección de proteínas virales y ácidos nucleicos	Detección rápida de virus conocidos con porciones invariables conocidas del genoma	Basado en amplificación	Detección rápida, barata y altamente sensible	Detección de ácido nucleico viral; resultados falsos positivos, dificultad para detectar virus con genomas muy variables; la detección de solo virus conocidos requiere el conocimiento de la secuencia de nucleótidos; para diferenciar entre virus infecciosos y no infecciosos
		Métodos relacionados con el uso de CRISPR/Cas	Se puede usar en el campo, tiempo relativamente corto para obtener resultados, alta sensibilidad	Detección de genoma viral; detección de solo virus conocidos, requiere el conocimiento de la secuencia de nucleótidos
		Hibridación sur/norte	Se pueden utilizar muchos tipos de muestras (sangre, líquido cefalorraquídeo, orina, lavado broncoalveolar, etc.)	Requiere conocimiento de la secuencia de nucleótidos del virus.
	Buscando nuevos virus	SNG	Determinación de la secuencia de nucleótidos de virus.	Dificultad para identificar virus de ARN en muestras de pacientes debido a etapas adicionales de preparación de la muestra y, como consecuencia, una disminución en la proporción de ARN viral a ARN del huésped. Dificultades en el procesamiento de datos
	Detección rápida de virus conocidos con proteínas virales conocidas	Inmunoensayos (determinación directa de proteínas virales y determinación indirecta de IgG e IgM)	Para las proteínas, la capacidad de detectar una exposición previa	Probabilidad de resultados falsos positivos; posible reactividad cruzada con virus estrechamente relacionados
Detección de partículas virales activas	Enfoque basado en células con detección de CPE	Enfoque basado en células con detección de EPC	Determinación de partículas virales en material clínico, estudiando su patogenicidad y mecanismos de transmisión. Detección de virus con un genoma muy variable	El principal problema es el largo período de tiempo (hasta varias semanas) requerido para que un resultado esté disponible. Los cultivos celulares también son muy susceptibles a la contaminación bacteriana y sustancias tóxicas en la muestra clínica de virus. Además, muchos virus no crecerán en cultivos celulares (virus de Epstein-Barr, hepatitis B, hepatitis C, parvovirus, etc.)
		Reportero basado en celdas		Es necesario desarrollar diferentes enfoques para estudiar virus específicos. Adecuado para virus con genomas anotados

Dolskiy, A. A.; Grishchenko, I. V.; Yudkin, D. V. Cell Cultures for Virology: Usability, Advantages, and Prospects. Int. J. Mol. Sci. 2020, 21 (21), 1–23. https://doi.org/10.3390/ijms21217978

2.1.2. Microscopía electrónica

El uso de la microscopía electrónica en un diagnóstico para la detección de un virus, suele ser muy útil, ya que es posible obtener un análisis rápido y preciso en una gran cantidad de moléculas, esta microscopía es sensible a partículas virales infecciosas y no infecciosas, sin embargo, no es un método convencional ya que suele ser demasiado costoso y lamentablemente no todos los laboratorios no tienen la disponibilidad de contar con este tipo de equipo, además para realizar la identificación más allá de la morfología se requiere de una prueba basada en anticuerpos para identificar el virus.

Dentro de la microscopía electrónica se encuentran varios tipos, como lo son la tinción negativa, siendo la más eficaz para la visualización de las partículas de virus debido a su simplicidad, rapidez y alta resolución. A su vez hay métodos para potenciar la visualización del virus, donde son la ultracentrifugación que se usa para concentrar las partículas de virus, la microscopía electrónica inmunológica, donde su uso se enfoca más en el inmunodiagnóstico rápido de la infección por el virus, aunque también se usa para potenciar la visualización de las partículas del virus, por último están las *"Thin Sectioning*", o Ultramicrotomía que a pesar de que sea un método menos rápido, da un testeo más confiable, específicamente cuando la estructura del virus no es distintiva por tinción negativa, aunque su mayor desventaja es que se necesita más tiempo para la preparación de la muestra y que se necesita a un personal capacitado.[12]

2.1.3. Microscopia de fluorescencia

En esta técnica la molécula fluorescente se une covalentemente a una inmunoglobulina que se produce contra una proteína específica algunas sustancias tienen la propiedad de luminiscencia. Emiten luz de un color cuando se exponen a la luz de un color diferente. Si la emisión de luz ocurre dentro de una millonésima de segundo de exposición a la luz, la luminiscencia es fluorescencia. Si la emisión de luz tarda más que esto, la luminiscencia es fosforescencia. El color de la luz emitida tiene una longitud de onda más larga que el color de la luz excitante.[13]

El desarrollo de microscopios para la detección de partículas virales de tamaño nanométrico hasta hoy en día se sigue perfeccionando, debido a las limitaciones

físicas en la óptica. Si bien las aplicaciones iniciales de la microscopía en virología se basaron predominantemente en microscopía electrónica, los campos de la biología celular y del desarrollo estuvieron fuertemente influenciados por la microscopía óptica. La microscopía óptica abarca modalidades de formación de imágenes que utilizan el espectro de la luz visible. Los productos químicos pequeños o las proteínas transgénicas expresadas en las células de interés en esta técnica son los fluoróforos que son los más utilizados.[14]

2.1.4. Serología

La serología se basa en la presencia de anticuerpos IgM específicos principalmente ya que también es afín a anticuerpos IgG. Dentro de los tipos de serología utilizados entran el ensayo neutralización donde generalmente se hacen diluciones con el virus y se incuban a óptimas incubaciones en placas de cultivo celular donde se monitorean para determinar el surgimiento de ECP, donde se tiene como control un anticuerpo monoclonal neutralizante al virus a estudiar, normalmente la multiplicación viral o ausencia de ECP indica la presencia de anticuerpos neutralizantes contra el virus, la dilución de suero se considera positiva a la presencia de anticuerpos neutralizantes si hay una reducción de la multiplicación viral mayor o igual al 50%.

Por otro lado, está la prueba de inhibición de la hemaglutinación, se emplea como un índice de reactividad para la evaluación de los niveles de anticuerpos, ha sido útil en la detección de antígenos; reemplazando las técnicas tradicionales y produciendo una mayor agilidad en la definición del diagnóstico, el ensayo de inhibición de la hemaglutinación de realiza una vez determinado el título hemaglutinante del virus a estudiar, normalmente se emplea una concentración constante del antígeno que se pone en contacto con concentraciones variables de los anticuerpos monoclonales neutralizantes o sueros específicos antes de adicionar la suspensión de eritrocitos.[15]

2.1.5. PCR (Reacción en Cadena de la Polimerasa)

La reacción de cadena de la polimerasa o PCR sirve para la detección de cantidades mínimas de ácidos nucleicos virales, debido a la amplificación exponencial de la secuencia diana. La PCR tiene como fundamento la medición continua en la suma de señales de fluorescencia en la reacción de amplificación. Esta técnica es un

método de amplificación de ADN *in vitro* mediante medios enzimáticos a tasas exponenciales en ciclos repetidos, bajo las siguientes etapas, la primera es un paso de desnaturalización del ADN de doble cadena extraído del material, la segunda es el paso de la hibridación agregando cebadores (oligonucleótidos), y la tercera y última es la reacción de extensión por el uso de una ADN polimerasa termoestable, agregando una enzima con el poder de sintetizar copias complementarias, la cantidad de ADN se duplica efectivamente a través de los pasos de extensión en cada ciclo de PCR.[16]

Dentro el mismo contexto, la cuantificación de secuencias diana mediante PCR en tiempo real o *real-time quantitative polymerase chain reaction (*RQ-PCR) tiene como fundamento la medición continua de la acumulación de señales de fluorescencia durante la reacción de acumulación, este método permite la detección del número de amplicones generados durante cada ciclo de amplificación de las muestras y permite el uso de sistemas de detección ya automatizados por completo, ya que además los resultados se muestran como gráficos de amplificación que resultan de una serie de mediciones de fluorescencia tomadas en puntos de tiempo definidos durante la amplificación, una de las características importantes de la tecnología en tiempo real es la capacidad de monitorear la cantidad creciente de producto en puntos de tiempo tempranos durante la reacción de PCR, la cuantificación mediante PCR en tiempo real no se ve afectada por concentraciones limitantes de reactivos ni por otras variables, como las condiciones de ciclo, que afectan la cuantificación en ensayos de PCR basados en análisis de punto final.

Las pruebas RQ-PCR optimizadas muestran una sensibilidad muy alta, con límites de detección entre 1 y 10 moléculas diana por reacción. Es gracias a esto que este método es hoy en día muy usado para detección de infecciones por virus, los formatos de detección basados en la hibridación específica de una o dos sondas de oligonucleótidos marcadas con fluorescencia con la secuencia objetivo durante la amplificación son los formatos informados con mayor frecuencia para la detección de virus en ensayos de diagnóstico. Dependiendo de la química utilizada, se han introducido diferentes tipos de sondas fluorogénicas, siendo la más común la sonda de hidrolisis.[17]

Las sondas de hidrólisis o sondas de nucleasa, son sondas de oligonucleótidos específicas del objetivo no extensibles que se unen a la hebra objetivo entre los cebadores de la PCR. Están doblemente marcados con un colorante informador fluorescente. Este principio físico se conoce como *transferencia de energía por resonancia de fluorescencia o FRET.*[17]

Este método se basa en el uso de dos sondas de oligonucleótidos que se hibridan una al lado de la otra con una secuencia ubicada entre los cebadores de amplificación. Las sondas están diseñadas para hibridar durante el paso de hibridación con la misma hebra en una disposición de cabeza a cola, a una distancia de 1 a 5 nucleótidos para acercar los dos tintes. El tinte donante es estimulado por una fuente de luz adecuada para emitir fluorescencia. Si ambas sondas se unen a las secuencias diana específicas, la energía de fluorescencia se transfiere de las moléculas donantes a las aceptoras (FRET) y el fluoróforo excitado emite una señal fluorescente, que se detecta y mide al final de cada paso de hibridación.[17]

2.1.6. Secuenciación de Sanger

La secuenciación de ADN hace posible precisar el ordenamiento de los nucleótidos del genoma de un organismo, microorganismo y entidad biológica. La secuenciación Sanger es un método de secuenciación enzimática, que determina la secuencia de un organismo mediante la terminación de la cadena a través de la incorporación de didesoxinucleótidos y hoy en día sigue siendo el método más utilizado.[18]

La tecnología de secuenciación de Sanger de ADN, ayudan a conocer la secuencia exacta de nucleótidos de cada genoma viral, por ello la secuenciación de ADN funge para la detección de virus no identificados, por igual al tener los datos genómicos del virus y su huésped, pueden ayudar a una pronta identificación de mutaciones importantes que le permiten al virus propagarse con facilidad, hay que tomar en cuenta que una mayoría de estos virus que causan brotes en la población son zoonóticos, esta secuenciación de ADN es un proceso que determina la secuencia de nucleótidos en un fragmento de ADN, esta secuencia en promedio se realiza con la tecnología del método de Sanger que necesita una plantilla de ADN monocatenario,

un cebador de ADN, una polimerasa de ADN, desoxinucleósidotrifosfatos normales (dNTP) y didesoxinucleósidotrifosfatos (ddNTP) (nucleótidos modificados).

La muestra de ADN viral es dividida en cuatro reacciones de secuenciación separadas y se agrega uno de los cuatro didesoxinucleótidos (ddATP, ddGTP, ddCTP o ddTTP), este método tiene como fundamento el uso de didesoxinucleótidos (ddNTP) para determinar el alargamiento de la cadena de ADN, el ADN viral que se secuenciará primero es desnaturalizado en cadenas simples mediante calor, después se reconoce un cebador con una de las hebras molde, dicho cebador es marcado de forma radiactiva o fluorescente que nos permitirá su detección del producto final en un gel de electroforesis, la clave de este método es que todas las reacciones parten del mismo nucleótido y terminan con una base específica.

Los secuenciadores de ADN separan las hebras en función de su tamaño mediante electroforesis capilar y detectan y registran la fluorescencia del colorante. Dado que los cuatro colorantes emiten fluorescencia a diferentes longitudes de onda, la identidad de cada banda se lee de acuerdo con las longitudes de onda en las que exhibe fluorescencia. Los datos obtenidos se muestran en forma de cromatograma y para la secuenciación del genoma completo de los virus de ARN, la estrategia más utilizada consiste en el diseño de amplicones superpuestos que abarcan todo el genoma, seguido de la amplificación dirigida de regiones genómicas mediante PCR de transcripción inversa. Los datos de secuencia generados luego se ensamblan.[16]

2.1.7. Secuenciación masiva

Teniendo como base el descubrimiento tradicional de virus por secuenciación de Sanger, posteriormente surge la secuenciación de alto rendimiento, por sus siglas en inglés *High throughput sequencing* (*HTS*) también llamada secuenciación de próxima generación o por sus siglas en inglés Next Generation Sequencing (NGS). Con la NGS se pueden caracterizar la naturaleza y composición de viromas completos.

La NGS tiene varios enfoques, desde la preparación de las bibliotecas genómicas hasta los métodos de preparación de muestras, como la digestión enzimática de ácidos nucleicos no virales y la exclusión por tamaño de los genomas no virales

mediante filtración en columna, ultrafiltración o centrifugación en gradiente de densidad. Dentro de estos destaca un enfoque de enriquecimiento de secuencias de virus llamado secuenciación de captura de viroma, centrado en la etapa de amplificación o preparación de la biblioteca genómica de la HTS, para aumentar la capacidad de caracterización de viroma. Este enfoque tiene el potencial para descubrir nuevos virus y para analizar de viroma, pero la complejidad de la técnica es muchas ocasiones es necesario hacer mejoras adicionales.[19]

La NGS ha demostrado su valía en la detección de virus y se ha convertido en una herramienta que ha reducido significativamente los costos de identificación. Sin embargo, esta técnica presenta un desafío debido a la variabilidad en el número de lecturas generadas para diferentes regiones genómicas o a una la cobertura desigual, que se puede generar.

En la mayoría de los casos, se observa una cobertura desigual del genoma viral en los datos de Whole Genome Sequencing y RNA-Seq, lo que afecta el porcentaje de cobertura que se puede lograr. Esto crea la necesidad de encontrar una cantidad óptima de datos para cubrir el genoma viral por completo o casi en su totalidad, sin generar en exceso datos de secuenciación.[16]

A pesar de estos desafíos, el uso de la secuenciación de próxima generación se ha establecido como un método para la detección de virus y para la metagenómica viral. Esto ha revolucionado la virología al permitir el descubrimiento de muchos virus novedosos.

Si bien la mayoría de los servicios comerciales de NGS ofrecen soporte bioinformático básico, como el ensamblaje de secuencias *de novo* o el mapeo a genomas de referencia, no suelen abordar los aspectos específicos de la detección y el descubrimiento de virus.[20]

Para abordar este desafío, se han desarrollado herramientas bioinformáticas específicas para la detección de virus humanos. Estas herramientas, por lo general, son paquetes independientes en línea de comandos de Unix que asignan las lecturas de NGS a genomas virales y realizan diferentes etapas de Blast para eliminar las lecturas que corresponden al hospedero. Aunque no existen programas de bioinformática que sean herramientas universales para el descubrimiento de virus, los

biólogos a menudo dependen de bioinformáticos profesionales para procesar los datos de NGS, lo que puede generar cuellos de botella en el análisis de datos.[20]

2.1.8. Generalidades de virus

En 1971, David Baltimore publicó en *Bacteriology Reviews* un artículo que hoy en día toma un papel vital en la virología que actualemente se titula B71, donde clasifica a los virus por las vías de transmisión de la información desde el ácido nucleico que se encuentra encapsulado en el virión hasta el ARNm, desde qué proteínas virales se traducen, además de que clasifica no solo los virus sino también las rutas de transferencia de información biológica.

Este sistema de Baltimore o B71 consta de seis clases de virus que se distinguen por sus distintas rutas de transferencia de información desde el ácido nucleico que se incorpora a viriones por lo que reflejan la naturaleza química y la polaridad del genoma, estas seis clases son las siguientes que se muestran en la tabla 3:[21]

Tabla 3. Clasificación de Baltimore.

TIPO	DESCRIPCIÓN
I	Virus de ADN de doble cadena (ds). Estos son los virus que encapsulan dsDNA y utilizan la ruta clásica de transmisión de información, la misma que en todas las células
II	Virus de ADN monocatenario (ss) que encapsulan ssDNA, que luego se replica y expresa a través de un intermediario de dsDNA
III	Virus dsRNA que empaquetan un genoma dsRNA que tiene que ser transcrito
IV	Virus de ARN de sentido positivo (+) que empaquetan en viriones un ssRNA de la misma polaridad que el ARNm para la síntesis de proteínas virales de modo que el ARN del genoma pueda traducirse directamente.
V	Virus de ARN (–) de sentido negativo que empaquetan un ARN que es complementario al ARNm y se transcribe para producir este último
VI	Virus de ARN de transcripción inversa que empaquetan un ARN de sentido positivo que se replica a través de un intermediario de ADN.

Koonin, E. V. The Baltimore Classi Fi Cation of Viruses 50 Years Later : How Does It Stand in the Light of Virus Evolution ? **2021**, *85* (3), 1–19.

2.1.9. Estructura y función

Los virus contienen un genoma de ARN o ADN rodeado por una capa protectora de proteína codificada. Para la propagación, los virus dependen de un huésped especializado, células que suministran la compleja maquinaria metabólica y biosintética de las células eucariotas o procariotas. Un virus completo en partícula se llama virión.

La función principal del virión es entregar su genoma de ADN o ARN a la célula huésped para que el genoma puede ser expresado por la célula huésped. El genoma

viral, a menudo con los elementos básicos asociados proteínas, sé empaqueta dentro de una cápside proteica simétrica. La proteína asociada al ácido nucleico, llamada nucleoproteína, junto con el genoma, forma la nucleocápside. En los virus envueltos, la nucleocápside está rodeada por una bicapa lipídica derivada de la membrana de la célula huésped modificada y tachonada con una capa externa de glicoproteínas de la envoltura del virus.[22]

2.1.10. Morfología

Los virus se agrupan según el tamaño y la forma, la composición química y la estructura del genoma, y modo de replicación. La morfología helicoidal se observa en las nucleocápsides de muchos virus filamentosos y pleomórficos. Las nucleocápsides helicoidales consisten en una matriz helicoidal de proteínas de la cápside (protómeros) envueltas alrededor de un filamento helicoidal de ácido nucleico. La morfología icosaédrica es característica de las nucleocápsides de muchos virus. El número y la disposición de los capsómeros (subunidades morfológicas del icosaedro) son útiles en la identificación y clasificación. Muchos virus también tienen una envoltura exterior.[22]

2.1.11. Composición química y modo de replicación

El genoma de un virus puede consistir en ADN o ARN, que puede ser monocatenario (ss) o bicatenario (ds), lineal o circular. El genoma completo puede ocupar una molécula de ácido nucleico (genoma monopartito) o varios segmentos de ácido nucleico (genoma multipartito). Los diferentes tipos de genoma necesitan diferentes estrategias de replicación.[22]

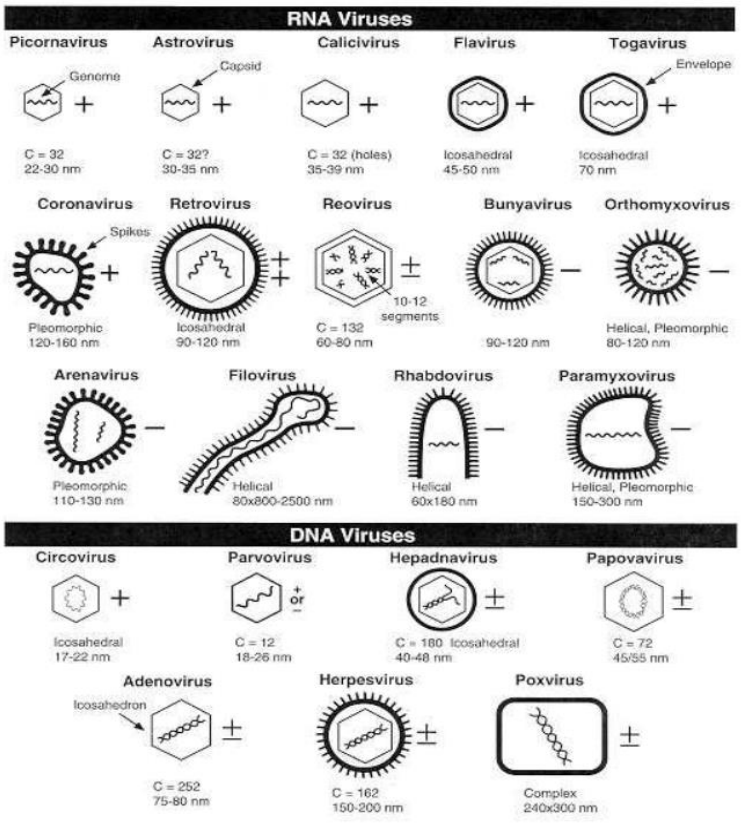
Los virus albergan una gran variedad de estructuras genómicas, siendo en su mayoría genomas de ARN, con casos especiales que algunos contienen genomas de ADN y algunos incluso de ADN y ARN, recordemos que varias clases de virus se clasifican según la polaridad de sus ARN, lo que indica si tiene información traducible (cadena positiva) o complementaria (cadena negativa).

2.1.12. Nomenclatura

Aparte de los datos físicos, la estructura del genoma y el modo de replicación son criterios aplicados en la clasificación y nomenclatura de los virus, incluida la composición química y la configuración del ácido nucleico, si el genoma es monopartita o multipartita. La hebra de ARN genómico de los virus de ARN monocatenario se denomina sentido (sentido positivo, sentido positivo) en orientación si puede servir como ARNm, y anti sentido (sentido negativo, sentido negativo) si es un complemento cadena sintetizada por una transcriptasa de ARN viral sirve como ARNm. También se considera en la clasificación viral el sitio de ensamblaje de la cápside y, en virus envueltos, el sitio de envolvimiento.[22]

Estos datos se han recopilado en diferentes bases de datos que son utilizadas por investigadores, y científicos para realizar estudios, investigaciones y análisis relacionados con los virus. A continuación, se describen las principales bases de datos en la figura 1 donde se esquematizan a los virus con diversos criterios distintivos como lo son presencia de una envoltura o (doble) cápside y genoma de ácido nucleico interno entre otros

Figura 1. Esquemas de diversos tipos de virus



Gelderblom, H. R. Structure and Classification of Viruses. Med. Microbiol. 1996, No. January 1996.

Capítulo 3. Bases de datos de virus

Una base de datos es una colección de información, conforme fue pasando el tiempo, se han logrado solucionar problemas para el almacenamiento, recuperación y manipulación de estos datos, esto con ayuda de un programa informático llamado sistema de gestión de bases de datos o DataBase Management System (DBMS), estos sistemas proporcionan cuatro ejes o elementos fundamentales:

- Código informático: Necesario para guiar al usuario a través del proceso de la base de datos
- Lenguaje informático: Para poder usar para agregar, manipular y consultar datos
- Herramientas: Hacen posible exportar los datos en una variedad de formatos
- Funciones administrativas: Necesarias para garantizar la integridad, seguridad y respaldo de los datos.[23]

En la actualidad, existen diversas bases de datos de virus disponibles que contienen información sobre diversas especies de virus. Estas bases de datos no solo almacenan información sobre cómo se clasifican los virus, sino que también incluyen otros aspectos importantes, como su estructura y función, morfología, composición química, modo de replicación y nomenclatura.

3.1. Comité internacional sobre la taxonomía de los virus (International Committe on Taxonomy of Viruses: ICTV)

El comité internacional sobre la taxonomía de los virus, o por sus siglas en inglés (Congress of the International Association of Microbiological Societies) (IAMS), es un comité de virología de la asociación internacional de sociedades de microbiología. El ICTV establece cada determinado tiempo una revisión, clasificación y actualización de la nomenclatura de los virus. Esto con la finalidad de establecer nuevos nombres taxonómicos y reconocer grupos naturales del virus.[24]

Para el estudio de cualquier entidad biológica nueva se necesita empezar con su clasificación y posteriormente nombrándolo, es aquí donde entra la taxonomía. La clasificación taxonómica es un esfuerzo científico mediante el cual los organismos

biológicos se agrupan y se colocan en una estructura jerárquica, en función de ciertas características evolutivas.

Los principios, procedimientos y nomenclatura utilizados para nombrar taxones están a cargo de una de las organizaciones internacionales encargadas de desarrollar las pautas necesarias. La taxonomía de los virus y demás es responsabilidad del ICTV que es el encargado de la tarea de desarrollar, refinar y mantener una taxonomía universal de virus por parte de la División de Virología de la Unión Internacional de Sociedades Microbiológicas. Hay seis subcomités, cada uno de los cuales cubre un grupo diferente de virus, diferenciados por el tipo de huésped que el virus infecta y la composición molecular del genoma del virus.[25]

3.2. Viralzone

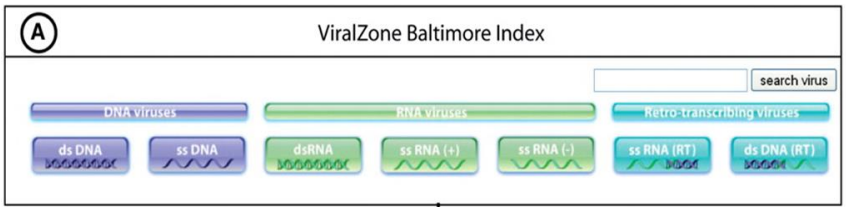
La interpretación de la diversidad molecular del virus es un tema de dificultad, por lo que es necesario un recurso integral que una el conocimiento de los datos teóricos de secuencias genómicas y proteómicas, como lo es la herramienta ViralZone. La base de datos ViralZone da acceso a hojas informativas sobre todas las familias/géneros de virus conocidos con un acceso fácil a los datos de secuencia, además de que ofrece ciertas herramientas visuales como imágenes de viriones detalladas y precisas. El propósito de ViralZone es vincular el conocimiento específico de cada familia del virus con secuencia genómicas y de proteínas virales, que es presentado en una hoja informativa donde se contiene información resumida sobre su genoma, ciclo de replicación, taxonomía y epidemiología.[26]

Una de las características destacadas de ViralZone es su índice taxonómico, que permite navegar y explorar los virus clasificados según la taxonomía de Baltimore. Esta clasificación se basa en las características estructurales y funcionales de los ácidos nucleicos de los virus, así como en su replicación y estrategias de transcripción. Esta taxonomía es ampliamente utilizada en el campo de la virología y proporciona una forma sistemática de clasificar y organizar los virus. En ViralZone, los usuarios pueden encontrar información detallada sobre cada género de virus, incluyendo características específicas, información sobre huéspedes infectados, datos epidemiológicos y más. Además, se proporcionan hojas informativas que contienen detalles sobre los virus y enlaces a recursos adicionales, a continuación,

se explicará la serie de pasos que hace esta herramienta siguiendo el orden de la figura 2-8:

(A) El índice ViralZone Baltimore proporciona información sobre la clasificación de virus según la taxonomía de Baltimore.

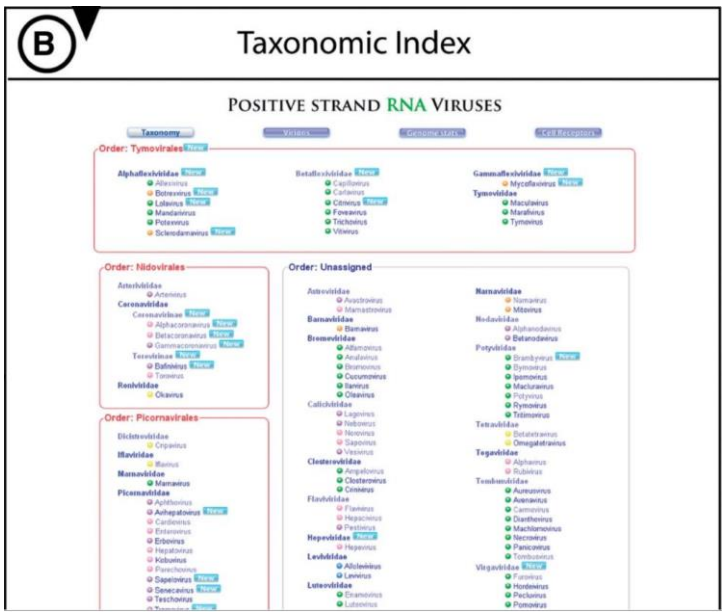
Figura 2. Esquematización A de la herramienta ViralZone



Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. Nucleic Acids Res. 2011, 39 (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901

(B) Se muestra un índice taxonómico de virus ssRNA (+) ordenados por orden, familia y género. Los colores indican los hospederos que puede infectar por cada género de virus: por ejemplo, rosa para humanos y otros vertebrados, púrpura para vertebrados no humanos, verde para plantas, amarillo para invertebrados, naranja para microorganismos del dominio eucariota y azul para procariotas

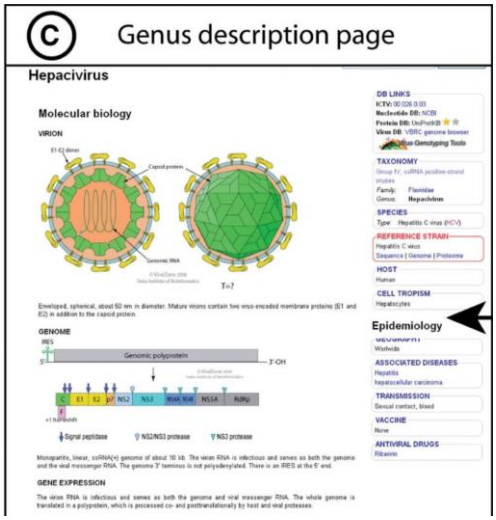
Figura 3. Esquematización B de la herramienta ViralZone



Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. Nucleic Acids Res. 2011, 39 (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901

(C) Cada género de virus tiene una hoja informativa que proporciona detalles específicos sobre ese género.

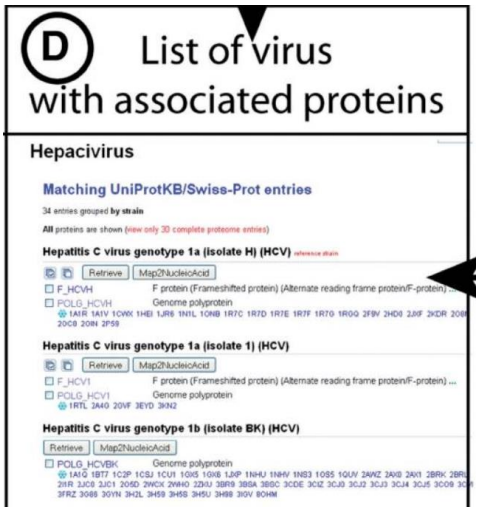
Figura 4. Esquematización C de la herramienta ViralZone



Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. Nucleic Acids Res. 2011, 39 (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901

(D) Se proporciona una lista de virus referenciados en UniProtKB/Swiss-Prot junto con las entradas de proteínas correspondientes que se muestran por defecto en la hoja de datos.

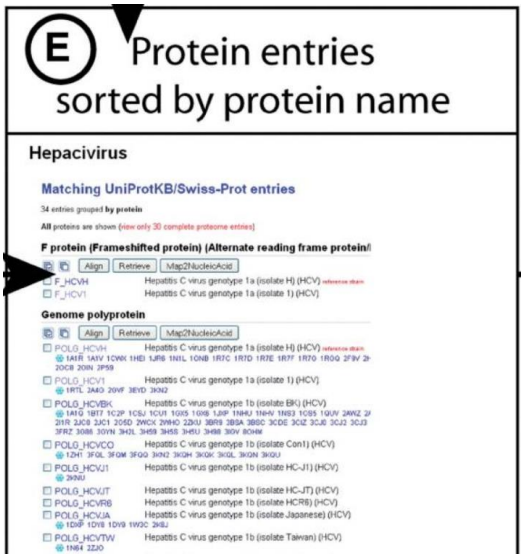
Figura 5. Esquematización D de la herramienta ViralZone



Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. Nucleic Acids Res. 2011, 39 (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901

(E) La lista de entradas está ordenada por nombres de proteínas.

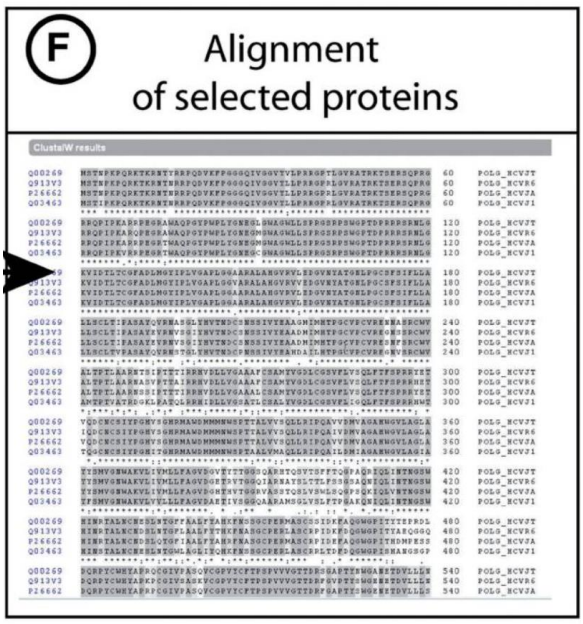
Figura 6. Esquematización E de la herramienta ViralZone



Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. Nucleic Acids Res. 2011, 39 (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901

(F) Después de seleccionar las entradas de proteínas en (E), se muestra una alineación correspondiente.

Figura 7. Esquematización F de la herramienta ViralZone



Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. Nucleic Acids Res. 2011, 39 (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901

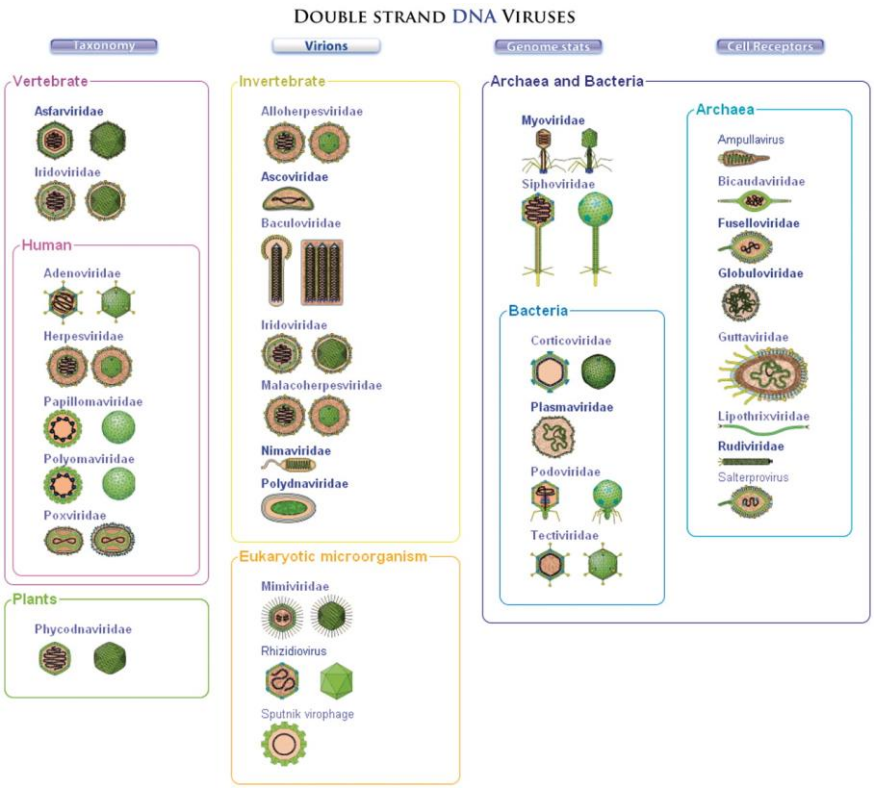
(G) Cada entrada de proteína Swiss-Prot contiene un enlace directo al sitio web de UniProt para acceder a detalles completos de la anotación de proteínas.

Figura 8. Esquematización G de la herramienta ViralZone

Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. *Nucleic Acids Res.* **2011**, *39* (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901.

La base de datos también ofrece enlaces a otras fuentes de información relevante, hacia las bases de datos UniProtKB/Swiss-Prot, donde se pueden encontrar detalles sobre las proteínas asociadas a los virus. Esto permite a los investigadores explorar la relación entre los virus y las proteínas que producen, así como acceder a información más detallada sobre la anotación de proteínas (Figura 9).

Figura 9. Esquematización de la herramienta ViralZone

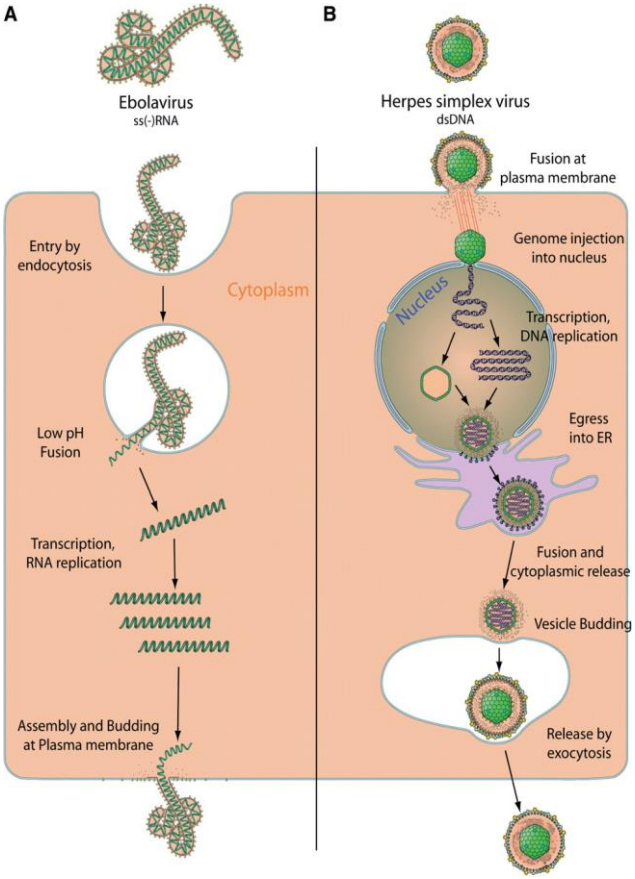


Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. *Nucleic Acids Res.* **2011**, *39* (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901.

Además, la página de ViralZone muestra una imagen pequeña del virión para todos los virus dsDNA en el enlace proporcionado. Al hacer clic en el nombre de la familia del virus o del género huérfano, se accede a una página de descripción del virus con una imagen del virión a tamaño completo.

En ViralZone, el punto de partida para acceder a las hojas de datos de virus son las siete páginas taxonómicas de acuerdo a la clasificación que ya se mencionó, que contienen la lista completa de familias y géneros de virus. Como por ejemplo en la figura 10 se esquematiza la ilustración los mecanismos de replicación viral del Ébolavirus y el Herpesvirus. La ventaja de esta base de datos es que se actualiza gradualmente, mientras que puede llevar años publicar nuevos libros con referencias de estos virus.[26]

Figura 10. Esquematización de los mecanismos del Ébolavirus y el Herpesvirus



Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. *Nucleic Acids Res.* **2011**, *39* (SUPPL. 1), 576–582. https://doi.org/10.1093/nar/gkq901.

3.3. Recurso de análisis y base de datos de patógenos de virus

La base de datos y recursos de análisis de patógenos de virus o por sus siglas en inglés Virus Pathogen Database and Analysis Resource (ViPR), da acceso a registros de secuencias, anotaciones de genes y proteínas, epítopos inmunitarios, estructuras 3D, datos de factores del huésped y otros tipos de datos a través de una interfaz de búsqueda intuitiva basada en la web. Los registros que son recuperados mediante consultas que son sometidos a diversos análisis como la alineación de secuencias múltiples, la inferencia filogenética, la determinación de variación de secuencias, la comparación mediante BLAST y el análisis estadístico de genómica comparativa basada en metadatos. Esta herramienta brinda apoyo a los investigadores en el campo de la virología que se dedican al estudio de agentes específicos y otros patógenos relevantes para la salud pública.[27]

La base de datos ViPR utiliza los genomas del Centro Nacional de Biotecnología e Información (NCBI) RefSeq para ampliar las anotaciones curadas manualmente de RefSeq al resto de los genomas que pertenecen al mismo taxón. También se utilizan métodos basados en secuencias para construir grupos homólogos de virus y las anotaciones asociadas, que se proporcionan en toda la herramienta para identificar fácilmente proteínas con funciones similares dentro de la familia de virus. En un esfuerzo por agregar información que no está incluida en los registros de GenBank, ViPR también ha curado manualmente la literatura científica para obtener información sobre el país, año y huésped de aislamiento de muchos virus clínicamente relevantes.

ViPR tienen la herramienta "Sequence Feature Variant Type" que registra la ubicación de regiones caracterizadas en las proteínas virales. Aunque esta funcionalidad se basa en trabajos anteriores realizados en proteínas HLA humanas e influenza se ha extendido para abarcar diversas taxonomías dentro de ViPR. Las definiciones de características de secuencia (SF) se obtienen de la literatura científica, registros de GenBank, UniProt y IEDB, y se clasifican en tipos como estructurales (por ejemplo, alfa-hélices), funcionales (por ejemplo, sitios activos), epítopos inmunitarios o alteraciones de secuencia.

Herramientas analíticas y de visualización integradas en ViPR:

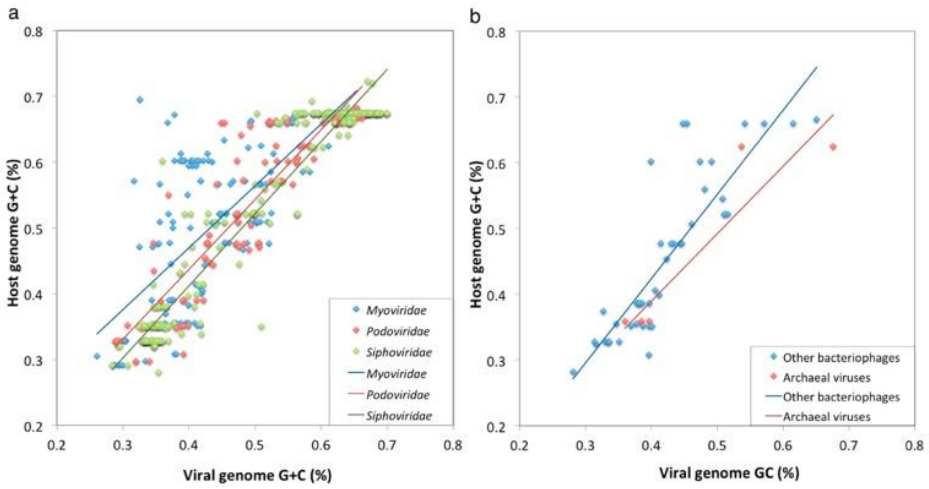
- MUSCLE: generación de alineaciones múltiples de secuencias (MSA) utilizando secuencias de nucleótidos o aminoácidos.
- ReadSeq: Convertir entre varios formatos de MSA.
- JalView: Visualizar y modificar alineaciones múltiples de secuencias de nucleótidos o aminoácidos.
- FastME, PhyML, RaxML: Inferir árboles filogenéticos para secuencias de nucleótidos o aminoácidos utilizando algoritmos de similitud o máxima verosimilitud.
- ModelCompare, ProtTest: Determinar qué modelo evolutivo usar al construir árboles de máxima verosimilitud.
- Archaeopteryx: Visualizar, manipular y decorar árboles filogenéticos.
- Jmol / Visualización: exploración de estructuras tridimensionales de proteínas.
- Meta-CATS: Comparar estadísticamente grupos de secuencias para identificar posiciones que difieren significativamente entre ellos.
- BLAST: Identificar secuencias de nucleótidos o aminoácidos mediante alineamientos en diversas bases de datos personalizadas de ViPR.
- Calculadora de Variación de Secuencias: Calcular la entropía presente en cada posición de nucleótidos o aminoácidos en grupos de secuencias de virus definidos por el usuario.
- Herramienta de Identificación de Péptidos Cortos: Encontrar cadenas cortas de aminoácidos en proteínas objetivo mediante coincidencia exacta, difusa o con patrones.
- Utilidad de Transferencia de Anotaciones Genómicas: Anotar una nueva secuencia de genoma utilizando un genoma de referencia bien anotado existente.
- Primer3: Diseñar cebadores de PCR para amplificar secuencias específicas de virus basadas en los datos dentro de ViPR.[27]

3.4. Virus-Host Data Base

La base de datos Virus-Host es de suma importancia para obtener información sobre el hospedero de los virus, ya que la replicación viral depende de los organismos hospederos. Tener acceso a información genómica y taxonómica de los virus, para

que posteriormente pueda ser posible correlacionar la composición de nucleótidos y codones en los genomas virales y poder conocer su coevolución y a su vez detectar interacciones genéticas, se logra mediante la base de datos Virus-Host. Esta se encarga de organizar enlaces con base en TaxID entre virus y sus hospederos, donde se extrae la información del huésped natural del genoma viral por medio de RefSeq y de las entradas de secuencias de proteínas por UniProtKB.[28] Para la predicción del hospedero, primero se examina el porcentaje y contenido genómico de guanina y citosina G+C como se puede ver en la imagen 11 donde se graficó el contenido G+C viral y genómico del huésped dado que los organismos huéspedes proporcionan una variedad de bloques de construcción moleculares y maquinaria necesaria para la replicación viral, las composiciones de nucleótidos de los virus.[28]

Figura 11. Contenido de G + C viral y genómico del huésped.



Mihara, T.; Nishimura, Y.; Shimizu, Y.; Nishiyama, H.; Yoshikawa, G.; Uehara, H.; Hingamp, P.; Goto, S.; Ogata, H. Linking Virus Genomes with Host Taxonomy. *Viruses* **2016**, *8* (3), 10–15. https://doi.org/10.3390/v8030066.

Ya con una información taxonómica realizada se evalúa un método computacional para la predicción del hospedero, con características genómicas de los virus y sus hospederos, se analiza la relación entre la similitud taxonómica del hospedero y la similitud de secuencia de genomas de los virus, esta suposición subyacente expone que si dos virus tienen genomas bastante similares en cuestión de secuencia y composición de nucleótidos, los dos virus pueden estar muy relacionados evolutivamente y pueden compartir el mismo hospedero.[28]

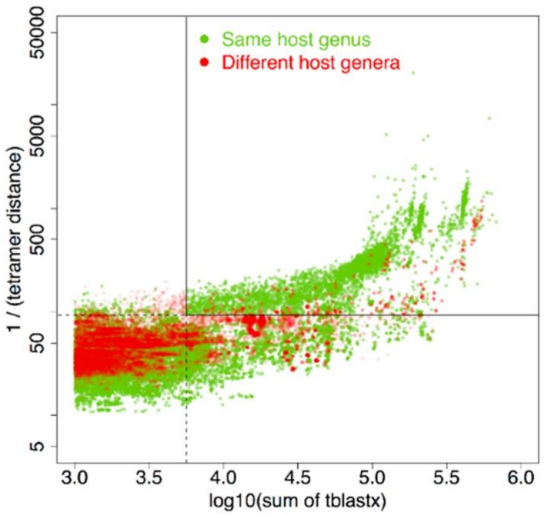
3.5 Viral Host Range database

La base de datos Viral Host Range (VHRdb) es una herramienta en línea para registrar, analizar y difundir interacciones de un rango de huéspedes con el virus. Es un esfuerzo de colaboración entre el Instituto Pasteur y la Universidad de Oxford, y está disponible gratuitamente para el público.

La VHRdb contiene información sobre más de 10 000 interacciones virus-huésped, incluido el nombre del virus, la especie huésped, el tipo de célula huésped y las referencias que respaldan la interacción.

La base de datos también incluye información sobre el rango de huéspedes de los virus, que es el rango de tipos de células y especies de huéspedes que un virus puede infectar, como se puede apreciar en la figura 12, donde se realizó la evaluación de la previsibilidad del rango de huéspedes basada en similitudes genómicas virales donde cada punto representa un par de genomas de virus y los colores de los puntos indican si los dos virus tienen el mismo huésped (verde) o no (rojo).

Figura 12. Evaluación de la previsibilidad del rango de huéspedes



Mihara, T.; Nishimura, Y.; Shimizu, Y.; Nishiyama, H.; Yoshikawa, G.; Uehara, H.; Hingamp, P.; Goto, S.; Ogata, H. Linking Virus Genomes with Host Taxonomy. *Viruses* **2016**, *8* (3), 10–15. https://doi.org/10.3390/v8030066.

La VHRdb es un recurso valioso para los científicos que estudian las interacciones virus-huésped. Se puede utilizar para identificar nuevas interacciones virus-huésped, para rastrear la evolución de los rangos de huéspedes del virus y para comprender

los factores que influyen en la especificidad del huésped del virus.[29]Estos son algunos de los beneficios de usar el VHRdb:

- Es una base de datos integral que contiene información sobre una amplia gama de interacciones virus-huésped.
- Está actualizado regularmente con nueva información.
- Está disponible gratuitamente para el público.
- Es fácil de usar y navegar.

3.6. Genomas Virales del Centro Nacional de Biotecnología e Información (NCBI)

Debido a la creciente cantidad de datos que se han ido resguardando hoy en día, son necesarios recursos de referencia que estén bien implementados para facilitar la identificación de secuencias ayudar en el ensamblaje de secuencias y a la vez fungir como fuentes de referencia. Es por ello que el recurso de genomas virales del NCBI funge como recurso de referencia, realizando un “shock wave” que mejora y facilita el uso de los datos de secuencias virales y ordenamiento de estas, cabe mencionar que también clasifica todas las secuencias genómicas de los virus disponibles públicamente y selecciona las secuencias genómicas de referencia.[30] Este acomodo y selección de genomas se basa en los criterios que son los siguientes

- Que todos los registros de RefSeq incluyan una notación de genes y proteínas, aunque sólo pueden incluir una anotación parcial
- Que la longitud del genoma sea válida de acuerdo a los estándares aceptados por la comunidad, por lo que la secuencia debe de cubrir toda la región de codificación del virus
- Que las secuencias de patentes y las sintéticas no se incluyen como genomas validados
- Que la creación de los registros de RefSeq para genomas virales por múltiples compuestos, esté representado por varios nucleótidos de RefSeq uno para cada segmento.

En la tabla 4 se muestran los genomas depositados en la base de datos de RefSeq del NCBI.[30]

Tabla 4. Enfoques actuales para la detección y caracterización de virus

Tipos de genoma	RefSeq Genomas Segmentados	Total de genomas	Total de secuencias INSDC
Virus dsDNA, sin intermediario replicativo	1 755	3 023	115 911
Virus dsRNA	919	17 929	56 699
Virus ssDNA	669	6 692	40 337
Virus de cadena negativa ssRNA	187	4 384	478 791
Virus de cadena positiva ssRNA, sin etapa de DNA	917	14 441	414 664
Virus de retro transcripción	123	8 614	727 762

Brister, J. R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Res.* **2015**, *43* (D1), D571–D577. https://doi.org/10.1093/nar/gku1207.

3.7. Base de datos de metagenómas virales IMG/VR

La base de datos IMG/VR se lanzó por primera vez en 2016 como un recurso independiente dedicado a los genomas virales dentro de la plataforma Microbiomas y genomas microbianos integrados o *Integrated Microbial Genomes & Microbiomes* (IMG por sus siglas en inglés). Mejoras experimentales y computacionales han ocasionado que el número de secuencias de genomas virales incrementen.

Debido a la expansión de secuencias virales metagenómicas, surge la necesidad de una plataforma computacional que anexe todas estas secuencias con metadatos asociados y herramientas analíticas, es por ello que IMG/VR es una base de datos pública con genomas de referencia y muestras metagenómicas. Las secuencias virales se pueden consultar utilizando una variedad de metadatos asociados, incluido el tipo de hábitat y la ubicación geográfica de las muestras, o la clasificación taxonómica según los genes virales distintivos.[31]

La creciente importancia de los genomas virales derivados de datos metagenómicos, también conocidos como "genomas de virus no cultivados" o "UViG", llevó al desarrollo de protocolos estándar y criterios de control de calidad para identificar, analizar y compartir mejor estos genomas.

Varios estudios han recopilado grandes colecciones de UViG, generalmente centrándose en un solo entorno o tipo de virus. Las secuencias que se importaron de RefSeq se asignaron a su taxonomía NCBI. UViGs derivados de los proyectos RVMT y GVMAG fueron designados a linajes taxonómicos listados en sus estudios

originales. Los UViG se asignan tentativamente a taxones virales, se utilizaron los siguientes métodos de clasificación taxonómica en orden de prioridad:

- Agrupamiento con genomas de virus RefSeq
- Asignación taxonómica basada en marcadores de geNomad
- Similitud con proteínas virales en NCBI NR
- Consenso vOTU

Los UViG se agruparon junto con las referencias RefSeq en vOTU recibieron la taxonomía del genoma de referencia como se puede. Para la asignación basada en marcadores, se empleó geNomad para clasificar secuencias utilizando perfiles de proteínas taxonómicamente informativos. Para asignar una taxonomía basada en la similitud con las proteínas virales del no-redundante NR de NCBI, MMseqs2 hace un módulo de taxonomía con los parámetros '- -start-sens 4 -s 6 - -sens-steps 2'. Finalmente, los UViG que no fueron designados a ningún taxón usando los métodos descritos anteriormente fueron asignados al taxón de consenso dentro de su vOTU, obtenido usando la función 'find_majority_vote' en taxopia. El taxdump de ICTV utilizado para todos los métodos de asignación taxonómica se generó utilizando TaxonKit.

Para la búsqueda del hospedero a los UViG identificados dentro de conjuntos de secuenciación del genoma completo aislados y genomas amplificados individuales (SAG), se les asignó una predicción de taxonomía del huésped basada en la taxonomía del genoma fuente. Los virus restantes se asignaron a huéspedes por asociación indirecta a través de coincidencias con una base de datos de espaciadores CRISPR o coincidencias *k-mer* exactas con genomas bacterianos y arquéales derivados de NCBI GenBank y estudios recientes de MAG a gran escala. Los espaciadores CRISPR se identificaron a partir de los 1,6 millones de genomas de los estudios NCBI y MAG usando una combinación de CRT y PILER-CR, finalmente cada virus se asignó al taxón huésped.

Además de la identificación de virus, los resultados de las anotaciones automáticas de geNomad se aprovecharon para proporcionar:

- Identificación automática de virus que probablemente empleen códigos genéticos alternativos

- Anotación a nivel de genes que se utilizan para encontrar virus que tienen una composición genética similar (ahora se utilizan en la herramienta 'Buscar UViG similares', como lo es Search and browse interface
- Asignación taxonómica usando un conjunto de perfiles taxonómicamente informativos
- Detección de características de virus, que se utilizan para identificar las secuencias que se asignaron a cada nivel.[32]

Un ejemplo, con el uso de Chinese Human Gut Viroma (CHGV), se obtuvieron 21,646 contigs virales no redundantes que fueron generados por ensamblajes combinados de lectura cortas (illumina) y largas (PacBio), posteriormente unas muestras clínicas se procesaron de acuerdo con un protocolo de enriquecimiento de viroma para obtener una gran cantidad de VLP, a continuación se extrajo ADN de cadena doble y alto peso molecular y se secuenciaron con ayuda de illumina HiSeq2000, las contaminaciones se eliminaron con conjuntos de datos vNGS y vTGS para identificar estos genomas virales se usaron las herramientas bioinformáticas: VirSorter v2.0, VirFinder v1.1 y PPR-Meta v1.1, además de que se usó una búsqueda BLAST contra los genomas Viral RefSeq usando BLASTn v.2.7.1.

Para saber si el virus era circular y cumplía con al menos dos siguientes criterios adoptados por gut virome database (GVD). Los parámetros que debe cumplir para asegurar que son genomas completos:

- Tener una puntuación en VirSorter ≥ 0,7,
- Tener un puntaje de VirFinder > 0.6,
- Tener puntaje de fagos en PPR-Meta > 0.7,
- Tener hits en Viral RefSeq con > 50 % de identidad y > 90 % de cobertura,
- Mínimo de tres ORF, que produzcan coincidencias BLAST en la base de datos NCBI POG 2013 con un valor E de ≤ 1e-5, con al menos dos por cada 10 kb de longitud de contig.
- Con la selección del catálogo CHGV usando CheckV, tener un criterio de >90% de integridad.

Alternativamente, los contigs cumplieron con uno de los criterios anteriores y fueron anotados como de alta calidad (≥ 90% de integridad) por CheckV y también fueron

anotados como virus. Los fagos de ADN bicatenario casi completos se seleccionaron del catálogo CHGV utilizando CheckV sobre la base de un criterio de selección de >90 % de integridad. Con fines comparativos, este subconjunto de genomas GPD se denominó GPD-HQ en este estudio.[32]

3.8. Base de datos de virus de plantas

Dentro del ámbito comercial, los virus también son una gran problemática para las plantas debido a que los estudios de estos son limitados y la diversidad de virus de plantas es muy basta. Los virus de plantas son de tipo de ADN o ARN y causan diversos tipos de enfermedades, algunas son fatales y pueden producir síntomas como son manchas anulares, patrones de mosaico, coloración amarillenta de las hojas y distorsión, agregando un crecimiento deformado. Hoy en día todavía no hay una cura efectiva para estas enfermedades virales en plantas, por lo que la mejor manera de controlar estos brotes es la destrucción completa de esta, lo que genera un daño al cultivo comercial. Muchos virus de plantas contienen el genoma de ARN monocateriano positivo y tienden a poseer genomas más pequeños que los virus de plantas de ADN. La mayoría de los viriones de plantas son pequeños debido al pequeño tamaño del genoma, aquí entran dos tipos de formas: filamentos y polígonos.

Existen también los viroides que son diminutas moléculas de ARN monocatenarias de unos pocos cientos de nucleótidos de largo y el virus satélite que son patógenos subvirares que dependen de su maquinaria de replicación. Las interacciones entre virus y plantas son complejas, ya que después de un tiempo los virus y plantas pueden coevolucionar; por un lado, los virus evolucionan pasando por altas tasas de mutación y por otro, las plantas pueden también mutar para combatir la infección por virus.

Debido al descubrimiento de una cantidad considerable de virus, se ha investigado la diversidad de estos virus gracias a la tecnología de secuenciación con estudios metagenómicos y de viroma, con ayuda de diversas herramientas con lo son:

- RNA-seq para identificar ciertos virus de ARN
- One Thousand Plant Transcriptomes Initiative que proporciono una cantidad enorme de datos de RNA-seq
- DPVweb, Plant Viruses Online, EPPO-Q-bank y base de daros de viroides.[33]

Capítulo 4. Herramientas bioinformáticas para la identificación de virus a partir de metagenomas

Debido a que hay una gran cantidad de virus sin cultivar, el uso de la metagenómica es una estrategia principal para el descubrimiento de nuevos virus, hasta hoy en día se han desarrollado muchísimas herramientas bioinformáticas de identificación de virus, esto a la vez dificulta la elección de herramientas, parámetros y los puntos de corte correctos, así que estas herramientas miden diferentes señales biológicas con diferentes algoritmos y bases de datos de referencia.[34]

4.1. Importancia de la identificación de virus en metagenomas

La NGS (Next Generation Sequencing) hace posible la evaluación de poblaciones microbianas complejas, con la ventaja de excluir el aislamiento y cultivo de cada especie microbiana, además de que no se requiere un conocimiento previo de secuencias microbianas de la muestra a analizar, la importancia de las aplicaciones de estas secuenciaciones son el descubrimiento de nuevos virus y la correcta y completa reconstrucción del genoma viral, ya sea a partir de preparaciones en cultivo o tomadas directamente de las muestras, es por ello que la NGS es muy viable como un método prometedor para el descubrimiento de virus emergentes con orígenes diversos.[35]

4.2. Desafíos en la identificación de virus en metagenomas

A pesar de que hoy en día las técnicas de secuenciación del genoma siguen creciendo a un ritmo drástico, los enfoques para reconstruir y clasificar los genomas virales a partir de muestras mixtas es un desafío a resolver, ya que afecta su rendimiento y utilidad, además de que sus desafíos técnicos han impedido su progreso, esto debido a que, por ejemplo, los fragmentos de los genomas virales suelen ser menos abundantes en órdenes de magnitud que los del hospedero, bacterias y/u otros organismos en metagenomas clínicos y ambientales, por lo que se han desarrollado programas para poder adaptarse a estos desafíos relacionados con la caracterización de secuencias virales.

Aunque ya sea posible una caracterización de entidades virales desconocidas con el uso de NGS, estas herramientas representan un cuello de botella práctico y

metodológico para un análisis eficaz, la mayoría de las herramientas bioinformáticas disponibles son complejas en su utilización para la mayoría de los usuarios principiantes, lo que exige experiencia y recursos informáticos de los que a menudo carecen los investigadores.[36]

4.3. Ensambles de genomas virales y recuperación a partir de ensambles

Hoy en día el uso de la metagenómica para descubrir y caracterizar poblaciones de microorganismos y virus en un nicho particular es cada vez más común. La identificación y el análisis de virus en entornos ambientales y de microbioma son relevantes para múltiples campos, en particular para la salud humana, se estima una abundancia enorme de partículas, los análisis metagenómicos han contribuido en gran medida a dilucidar la verdadera diversidad de los virus, sin embargo, los estudios metagenómicos individuales pueden identificar potencialmente hasta varios miles de genomas de fagos nuevos no clasificados, y los que no pueden clasificarse o anotarse completamente permanecen como materia oscura viral (Viral dark matter), la mayoría de estos virus no se pueden cultivar, y los genomas virales no cultivados (Uncultivated Virus Genome UViG) representan la diversidad actual en las bases de datos públicas.

Dadas las diferencias en las secuencias genómicas virales, tanto entre distintos tipos de virus como entre virus y organismos celulares, la metagenómica viral plantea múltiples desafíos, esto ocasiona que se ha desarrollado una sobreabundancia de herramientas y flujos de trabajo metagenómicas para manejar todos los aspectos de estos análisis, algunas de ellas especializadas para su uso en fagos, virus de eucariontes, y otras diseñadas para todos los datos virales, esto dificulta mantenerse al día con las ofertas disponibles y seleccionar una herramienta específica con parámetros adecuados para un conjunto específico de datos metagenómicos, por lo que es necesario enfocarse principalmente en cinco temas:[37]

- Ensamblaje e identificación de secuencias de virus (Virus de eucariontes, Bacteriófagos, profagos)
- Anotación del genoma de virus.
- Clasificación taxonómica de los genomas virales. Análisis de interacción virus-huésped.
- Diversidad de virus.

Las infecciones virales siempre han sido un problema para las poblaciones, por lo que el uso de los enfoques metagenómicos es utilizado más frecuentemente que antes en la detección de patógenos virales emergentes, a la vez de la detección de nuevos patógenos virales.

Es así como la identificación *in silico* de genomas virales completos a partir de datos de secuencia haría posible una eficaz y oportuna caracterización filogenética de estos emergentes virus. El ensamblaje de genomas virales individuales a partir de un metagenoma es un desafío, no solo por la falta de un genoma de referencia, sino también por la variación y la cobertura desigual o insuficiente.

Las estrategias metagenómicas para el descubrimiento de virus se basan en la amplificación de ácidos nucleicos independiente, generando así lecturas (regiones pequeñas de secuencias del genoma). Los métodos comunes de amplificación aleatoria son la amplificación por desplazamiento múltiple (MDA) o la amplificación con un solo cebador independiente de la secuencia (SISPA). Las ventajas de la amplificación independiente de la secuencia son la simplicidad y la velocidad, más la capacidad de identificar y secuenciar cientos de virus simultáneamente, lo que permite la detección de virus nuevos o no reportados previamente, diferentes de los ya descritos. Inherente al enfoque es que una gran fracción del metagenoma consiste en secuencias de otros organismos, además de los objetivos virales, incluidas secuencias del huésped, arqueas y bacterias, a pesar de las estrategias de enriquecimiento físico para partículas de virus que se aplican a menudo.[38]

Después de realizar la secuenciación de los genomas virales por los métodos anteriores descritos, el proceso que sigue ahora es el ensamble del genoma. Si la secuencia de referencia difiere significativamente de la muestra o si no hay una referencia disponible, es necesario generar una secuencia de consenso mediante el ensamblaje *de novo* de las lecturas. Es decir, se tiene que hacer un ensamble de novo para tener todos los genomas virales recuperados de los metagenomas.

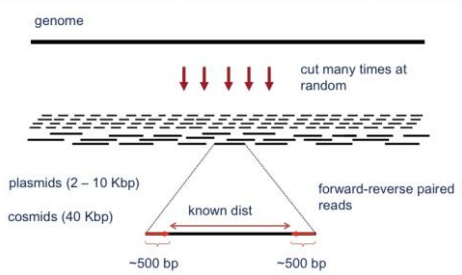
Un factor a tomar en cuenta al momento del ensamble de los genomas, es determinar la calidad de las lecturas, por lo que es fundamental recortar las bases de mala calidad y eliminar las secuencias de adaptadores de los conjuntos de datos de lectura antes

de ejecutar los ensamblajes. La mayoría de los algoritmos utilizados para el ensamblaje *de novo* se dividen en dos grupos:

- Ensambladores de consenso de diseño superpuesto (OLC)
- Ensambladores de gráficos de Bruijn.

Los ensambladores de OLC, como MIRA, Edena, Canu, y Newbler primero identifican pares de lecturas que se superponen y luego construyen un gráfico en el que las lecturas están representadas por nodos en el gráfico, con lecturas superpuestas conectadas por bordes. Luego, el gráfico se analiza para encontrar rutas a través del gráfico que atraviesen múltiples bordes, lo que permite que las lecturas se coloquen en mosaico en el orden correcto para generar una secuencia del genoma. Sin embargo, el enfoque OLC normalmente no escala bien porque el gráfico de superposición puede llegar a ser muy grande y que busque todos los sobrelapes posibles sin tomar atajos, lo que los hace computacionalmente exhaustivos, esto se puede ver claramente en la imagen 13.

Figura 13. Composición de IMG/VR v4 con respecto al origen de UViG



Otros ensambladores, como ABySS, Velvet, SPADES y A5 utilizan un algoritmo gráfico de Bruijn, que reduce el esfuerzo computacional al dividir las lecturas en cadenas más cortas de una longitud fija *k* (llamadas *k-meros*). El gráfico de Bruijn captura superposiciones de longitud *k-1* entre estos *k-meros*, lo que evita la necesidad de calcular superposiciones entre secuencias largas. Las lecturas en sí mismas no se modelan, sino que se representan mediante rutas a través de un gráfico; por lo tanto, este enfoque ha demostrado ser muy eficaz y utilizar menos memoria computacional. Sin embargo, existe un límite superior para la longitud de *k*, los ensamblajes *de novo* suelen consistir en varias secuencias largas y contiguas (contigs) en lugar de

genomas completos, los errores de secuenciación, las regiones repetidas y las áreas con poca cobertura tienden a confundir el proceso de ensamblaje.

Los contigs se pueden unir para producir un genoma draft (borrador de genoma) mediante la alineación con una secuencia de referencia viral relacionada utilizando un software como Abacas, Scaffold Builder y CONTIGuator. Si no hay una referencia disponible, se pueden utilizar lecturas de pares de extremos o lecturas de pares para armar los contigs en el orden lineal correcto y producir un borrador del genoma que contiene lagunas. Muchos paquetes de ensamblaje *de novo* llevan a cabo este paso de andamiaje automáticamente en contigs ensamblados cuando se les dan datos de lectura de extremos emparejados, pero también hay disponibles andamios independientes, incluidos Bambus2 y BESST. Como alternativa, los ensamblajes realizados con datos de lecturas cortas, como los de las plataformas Illumina o Ion Torrent, se pueden mejorar mediante el uso de un segundo conjunto de datos con lecturas más largas para unir contigs.

El software que rellene los huecos o en inglés “gaps” puede usarse para cerrar algunas partes del genoma sin ensamblar las que quedan en el borrador del genoma. Sus algoritmos requieren datos de extremos emparejados e identifican pares de lectura específicos en los que un miembro coincide con el final de un contig y el otro cae dentro del espacio. Dichos subconjuntos de lectura se utilizan para extender los contigs de forma iterativa y cerrar las brechas mediante la superposición de *k-mer* o el ensamblaje local de novo.

Sin embargo, las regiones repetidas con un período más largo que la longitud de lectura no se pueden resolver; estos requieren PCR seguida de secuenciación de Sanger o datos de una plataforma HTS que produce lecturas más largas, se han desarrollado paquetes como ICORN2 para automatizar la verificación y corrección de errores de los genomas virales, mientras que algunos ensambladores son capaces de llevar a cabo la mayoría de los procesos descritos en esta sección.[38]

La elección del programa de ensamblaje para datos metagenómicos virales es importante para la identificación precisa de contigs virales y análisis posteriores. En un estudio reciente se revisaron la eficacia de 16 ensambladores de lectura corta, especializados o no en datos metagenómicos. Recomiendan el uso de

MetaviralSPAdes, para el ensamble del viroma, seguido de MEGAHIT. La presencia de secuencias repetidas, así como valores de cobertura demasiado altos y demasiado bajos, son los principales obstáculos para el ensamblaje eficiente de los datos de viroma.

MetalviralSPAdes es una herramienta que modifica varias herramientas de metaplasmidSPAdes para hacer posible la secuenciación viral y se basa en una serie de tres pasos:[39]

- Ensamblaje viral

Que encuentra subgráficos virales putativos en un gráfico de ensamblaje metagenómico y genera contigs en estos gráficos

- Verificación viral

Que es para verificar si los contigs resultantes tienen origen viral

- Complete viral

Para verificar si estos contigs representan genomas virales completos

El programa ViralAssembly es una adaptación de MetaSPAdes para el ensamblaje de datos virales. Aprovecha la detección de secuencias genómicas circulares de MetaplasmidSPAdes para detectar genomas virales circulares y permite la detección y el ensamblaje de repeticiones terminales en genomas lineales, como se aprecia en las imágenes 9,10 y 11.

En un análisis de 18 conjuntos de datos reales de virome, se demostró que ViralAssembly supera a MetaSPAdes en términos de integridad de contig. MetaFlye es un ensamblador viral no específico que ha demostrado detectar y ensamblar genomas virales en conjuntos de datos metagenómicos de lectura larga con buena eficiencia. Mientras que el programa VirION, presenta una segunda iteración, y emplea lecturas cortas para corregir errores de secuenciación en ensamblajes de lectura larga y supera a los ensambladores híbridos y de lectura corta cuando se prueba en viromas de ADN de doble cadena (dsDNA). Los tres programas recomendados, MetaviralSPAdes, metaFlye y VirION2, están disponibles para descargar desde GitHub. Cada uno requiere un conocimiento básico del uso de la línea de comandos para la instalación, así como la instalación de ciertas dependencias.[39]

4.4. Herramientas basadas en homología

La metagenómica hoy en día es un método muy usado para el descubrimiento de virus, ya que la mayoría permanecen sin cultivar, sin embargo, el hecho de detectar virus en datos metagenómicos no es algo común, es por ello que recientemente están al alcance varias herramientas bioinformáticas de identificación de virus. Esto a la vez dificulta la elección de dichas herramientas, debido a que esta gran gamma de herramientas usas diferentes metodologías.

Es de vital importancia realizar una comparación independiente para brindar una guía objetiva a los operadores; esta evaluación comparativa brinda orientación sobre las opciones de herramientas bioinformáticas de identificación de virus y brinda sugerencias para ajustes de parámetros para investigadores de virómica.[35]

Los métodos basados en homología buscan virus mediante la alineación de secuencias de contigs con bases de datos de genes virales. Por lo general, esta comparación se realiza utilizando la búsqueda BLAST contra una base de datos.

En algunos casos, se realiza la comparación utilizando secuencias de aminoácidos, ya que son más conservadas evolutivamente. Algunas de las herramientas más populares se centran en la detección de virus presentes en el entorno, como VirSorter[40], que utiliza características distintivas de los virus para su identificación.

Por ejemplo, VirSorter se alinea con genes virales "distintivos" y realiza la búsqueda basándose en la similitud utilizando BLASTP y HMM en una base de datos previamente construida.[40] Con este programa es posible identificar nuevos virus, aunque se enfoca principalmente en la búsqueda de bacteriófagos.

Dentro de esta categoría se encuentran los servidores web y las herramientas web. Estas herramientas web hacen que el proceso sea más accesible para personas con pocos conocimientos de bioinformática. Algunos ejemplos de estos servidores web son VirAmp[41], Metavir2[40] y Viral Informatics Resource for Metagenome Exploration (VIROME)[42], que reciben contigs como datos de entrada.

Estas herramientas están diseñadas para buscar virus ambientales, especialmente bacteriófagos. Además, hay herramientas web como Phaster[43] que comparan

secuencias de consulta con una base de datos de genes virales. También se pueden utilizar herramientas bioinformáticas para realizar una clasificación taxonómica en metagenomas, lo que podría ayudar a identificar secuencias virales. Estas herramientas no están especializadas exclusivamente en la búsqueda de virus, pero pueden resultar útiles para este propósito.

Cabe destacar que en los virus la mayoría de las señales permanecen invisibles en conjuntos de secuencias metagenómicas, debido a la ausencia de marcadores genéticos universales, representantes de bases datos y herramientas de identificación obsoletas

Es por ello que VirSorter2, una herramienta que identifica los virus de ADN y ARN aprovecha los avances de la base de datos informada por el genoma mediante el uso de una colección de clasificadores automáticos personalizados con el fin de aumentar la precisión y el rango de detección de secuencias de virus, además esta herramienta fue muy eficaz a la hora de minimizar errores asociados con las secuencias genómicas atípicas, como las islas de patogenicidad y/o plásmidos.

El diseño modular de VirSorter2 lo hace intrínsecamente capaz de expandirse a nuevos tipos de virus a través del diseño de nuevos clasificadores para mantener la máxima sensibilidad y especificidad.[44]

La introducción de la metagenómica viral hizo posible una mejor compresión de la diversidad genética que abarcan los genomas de los fagos y gracias a la amplia gama de biomas de genomas de fagos y que va en forma creciente ha impulsado el desarrollo de herramientas computacionales que caracterizan el multinivel de nuevos fagos, gracias a que se basa solo en secuencias genómicas, taxonómicos, como se puede apreciar en la tabla 5 donde se describe las herramientas de detección de fagos. Aunque las herramientas disponibles han contribuido enormemente a nuestro conocimiento de la diversidad y la ecología de los fagos, el aumento continuo de los programas de software hace que sea un desafío mantenerse al día con ellos y con el propósito para el que cada uno está diseñado.[22]

Tabla 5. Descripción general de las herramientas de detección de fagos metagenómicos publicadas

Herramienta y versión	Categoría	Método	Conjunto de entrenamiento/Base de datos de referencia	Distribución
DeepVir Finder (1.0)	Secuencia	Red neuronal de aprendizaje profundo basada en *k -mer*	Genomas NCBI RefSeq antes de mayo de 2015 y secuencias de viroma	GitHub
MARVEL (0.2)	Homología	Bosque aleatorio que utiliza densidad de genes, cambios de cadena y proteínas	Genomas NCBI RefSeq antes de 2016	GitHub
MetaPhinder	Homología	Análisis integrado de aciertos de BLASTn en una base de datos de fagos	Conjunto de datos virales de genomas NCBI, genomas EMBL EBI, phageDB, PhAnToMe/conjunto de datos bacterianos de genomas NCBI. Descargado antes de agosto de 2014	GitHub
Phamers	Secuencia	*k* - Vecinos más cercanos y métrica de proximidad centroide de *k* mers	Genomas NCBI RefSeq antes de octubre de 2015	-
PPR-Meta	Secuencia	Red neuronal convolucional (CNN) de codificaciones one-hot de nucleobases y codones	Genomas NCBI RefSeq. Fecha de descarga desconocida.	-
RNN-VirSeeker	Secuencia	Memoria a corto plazo (LSTM) de secuencias	Genomas NCBI RefSeq descargados antes de enero de 2014	-
Seeker	Secuencia	LSTM de secuencias	Genomas NCBI y genomas EMBL EBI. Fecha de descarga desconocida.	PyPi
Unlimited breadsticks	Homología	HMM de los genes característicos del virus	–	GitHub
Vibrant (1.0.1)	Homología	Red neuronal de firmas de proteínas que incluyen proporciones de aciertos de KEGG, VOG y PFAM, y la presencia de genes clave de tipo viral.	NCBI RefSeq y Genbank antes de julio de 2019	Bioconda
ViralVerify (1.1)	Homología	Clasificador Naive Bayes usando una hmmsearch de genes predichos con Prodigal	Genomas NCBI RefSeq. Fecha de descarga desconocida.	Bioconda
ViraMiner	Secuencia	CNN de nucleobases codificadas en caliente	Secuencias de 19 metagenomas WGS de muestras de microbioma humano.	–
VirFinder (1.1)	Secuencia	Regresión logística usando *k* -mers	Genomas NCBI RefSeq descargados antes de enero de 2014.	Bioconda
VirMine	Homología	Búsqueda BLAST de ORFs contra bases de datos virales y no virales	Conjunto de datos virales: genomas virales NCBI RefSeq. Conjunto de datos bacterianos: COG bacterianos. Fecha de descarga desconocida.	–
VirMiner	Homología	Bosque aleatorio (RF) basado en perfiles funcionales y homología de proteínas	Conjunto de datos virales Genomas NCBI y base de datos ACLAME. Conjunto de datos bacterianos: genomas NCBI. Descargado en octubre de 2016.	–
VirNet	Secuencia	Modelo de atención profunda de secuencias.	Genomas NCBI RefSeq. Fecha de descarga desconocida.	–
VIROME	Homología	Información funcional y taxonómica basada en homología ORF	–	–
VirSorter	Homología	Homología de genes, incluido el enriquecimiento de genes cortos y de tipo viral, y el agotamiento de los aciertos de PFAM y los cambios de cadena	Genomas NCBI RefSeq antes de enero de 2014 y viromas ambientales.	wget
VirSorter2 (2.2)	Homología	Clasificadores aleatorios de bosques usando una hmmsearch de genes predichos con Prodigal	Genomas NCBI RefSeq antes de enero de 2020 y genomas de alta calidad de la literatura.	Bioconda
Buscador de virus	Homología	Búsqueda BLAST en la base de datos de virus, seguida de búsqueda en la base de datos completa del NCBI para eliminar los falsos positivos	Base de datos NCBI NT y NR solo viral antes de agosto de 2016	–

Gelderblom, H. R. Structure and Classification of Viruses. *Med. Microbiol.* **1996**, No. January 1996.

4.5. Herramientas basadas en máquinas de aprendizaje

En la clasificación de virus, se emplean las arquitecturas de aprendizaje automático como otra estrategia. Estas arquitecturas requieren una secuencia de entrada que esté clasificada según características específicas. Los algoritmos de

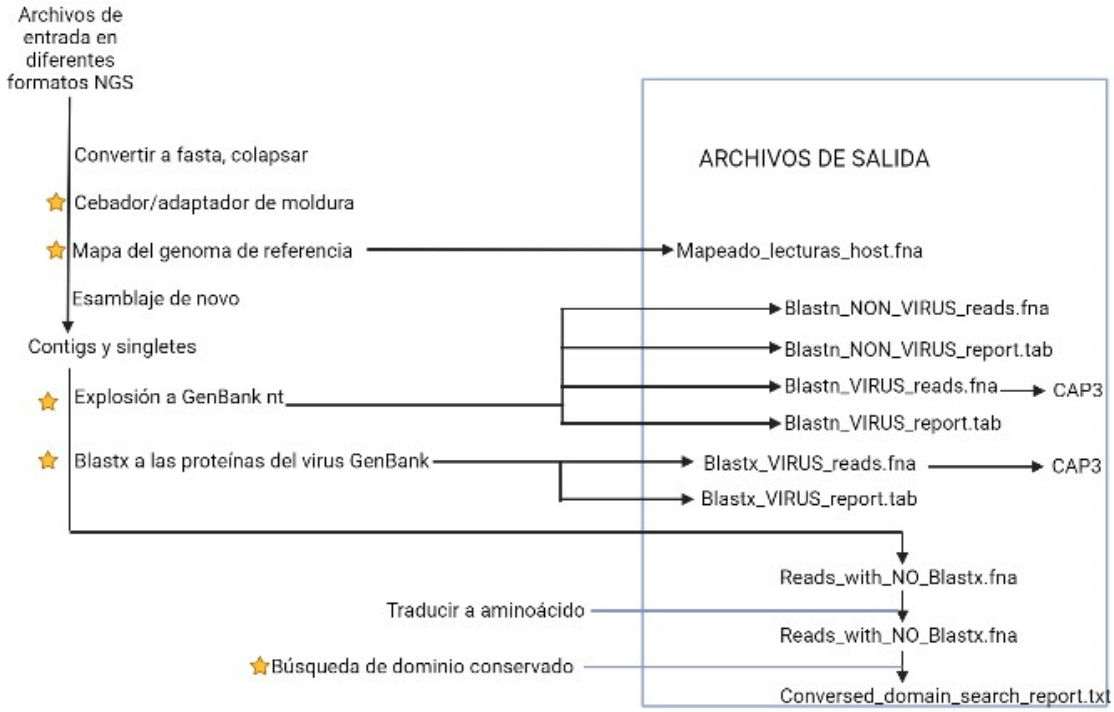
aprendizaje automático nos permiten identificar patrones o características en las secuencias y hacer predicciones utilizando enfoques como k-mer, aprendizaje automático de bosque aleatorio y redes neuronales convolucionales entrenadas.[45]

Virfinder[45] es uno de los programas más populares que utiliza este enfoque. Emplea el aprendizaje automático para identificar firmas virales en contigs, y ha demostrado tener una mejor capacidad de identificación de genomas virales en comparación con VirSorter. Sin embargo, es importante tener en cuenta que Virfinder es un modelo clasificador entrenado con secuencias de la base de datos RefSeq, lo que limita su capacidad para identificar ciertos grupos virales. Por ejemplo, si se le proporcionan virus representativos de ambientes marinos, no será capaz de identificar adecuadamente virus de otros ecosistemas. Por lo tanto, es necesario contar con un amplio conjunto de viromas y genomas virales para obtener datos de entrenamiento más completos y así encontrar nuevos virus.[45]

VirFinder, se basa en la frecuencia *de k*-mer para la identificación del código del virus que evita por completo las búsquedas de similitudes basadas en genes, esta herramienta identifica las secuencias virales en función de la observación empírica de que los virus y los huéspedes tienen firmas *k*-mer perceptiblemente diferentes, es por ello que esta herramienta basada en *k*-mer complementa los enfoques basados en genes y mejora significativamente la identificación de secuencias virales procarióticas, especialmente para estudios metagenómicos de ecología viral.[46]

Los métodos existentes basados en referencias y basados en homología de genes no son eficaces para identificar virus desconocidos o secuencias virales cortas a partir de datos metagenómicos, es por ello que hay un método de aprendizaje automático sin referencias ni alineaciones, como DeepVirFinder que se observa en la imagen 14, sirve para identificar secuencias virales en datos metagenómicos mediante el aprendizaje profundo, esta ampliación de los datos de entrenamiento con millones adicionales de secuencias virales purificadas de muestras de metaviroma mejorando la precisión para identificar grupos de virus que están subrepresentados.[47]

Figura 14 Diagrama de flujo de VirFind



Ho, T.; Tzanetakis, I. E. Development of a Virus Detection and Discovery Pipeline Using next Generation Sequencing. **2014**, *473*, 54–60.

4.6. Tuberías (*Pipelines)* para la identificación de virus

Las tuberías o *pipelines* (del inglés) de análisis para recuperar genomas virales se centran en virus que infectan a los humanos y se obtienen a partir de muestras ambientales. Estas tienen como objetivo realizar análisis a partir de datos sin procesar, sin ningún pretratamiento o alineación de lecturas cortas con una base de datos de referencia.

Por lo general, las tuberías comienzan el análisis a partir de datos brutos de secuenciación de nueva generación (NGS) utilizando control de calidad, luego realizan recortes de secuencias, eliminan lecturas cortas y retiran secuencias de adaptadores. Algunas herramientas mapean o alinean las lecturas con genomas de huéspedes de referencia para eliminar la contaminación del huésped.

Algunos flujos de trabajo eliminan secuencias duplicadas (de duplicación) y nucleótidos no resueltos (Ns), y posteriormente realizan un ensamblaje de las lecturas para realizar una clasificación taxonómica de acuerdo con una base de datos como NCBI, y los virus se clasifican según los criterios del Comité Internacional de

Taxonomía de Virus (ICTV). Sin embargo, cada tubería utiliza estrategias particulares para recuperar genomas virales dentro de un metagenoma viral, como en la tabla 5 donde se visualizan las bases de datos de proteínas. El análisis de las secuencias bioinformáticas varía según el tipo de muestra que se está analizando; en el caso de trabajar con un viroma humano u otro huésped, las secuencias correspondientes deben eliminarse, mientras que en muestras de origen ambiental no es necesario eliminar al huésped.

Muchas de estas tuberías están diseñadas para buscar virus que causan enfermedades en humanos, por ejemplo, VirusHunter[48], Sequence-Based Ultrarapid Pathogen Identification (SURPI)[49], Virus Identification Pipeline (VIP)[50], VirusSeeker[51], ViromeScan[52], Viral Genome-Targeted Assembly Pipeline (VirusTAP)[53], virMine[54], DisCVR[55] y VirAnnotOTU[56].

Estos programas suelen mapear las lecturas con el genoma humano, y luego las lecturas filtradas se utilizan en análisis posteriores. Para los virus ambientales, existen VirAmp[41], Holovir[57] y FastViromeExplorer[58], mientras que otros programas se centran en la recuperación de virus de ARN, como short RNA subtraction and assembly (SRSA), y otros se enfocan en virus de ARN en plantas, como VirFind[20], Kodoja[59] y VirusDetect[60]. Sin embargo, estos programas tienen enfoques diferentes; por ejemplo, VirFind utiliza BLAST y Kodoja utiliza la estrategia de análisis de k-mer a partir de datos de RNA-seq. VirusDetect se utiliza para buscar virus mediante el ensamblaje de pequeños ARN totales[60].

Muchos de estos programas identifican virus, pero pocos realizan una clasificación taxonómica; algunos ejemplos son VirAnnotOTU, VIP, SRSA, HoloVir y ViromeScan. Para aquellos que no realizan esta asignación taxonómica, es necesario utilizar otras herramientas, las cuales se discutirán más adelante.

La herramienta informática LAZYPIPE identifica virus ya conocidos y nuevos ubicados en muestras ambientales, LAZYPIPE es un pipeline que se basa en Unix para el ensamblaje automatizado y creación de perfiles taxonómicos en bibliotecas NGS en forma de scripts C++, Perl y R. LAZYPIPE tiene ciertas ventajas para la elaboración de perfiles taxonómicos de datos NGS virales, esta herramienta subcontrata la

búsqueda de homología a un servidor independiente, lo que suprime la necesidad de instalar y actualizar continuamente estas bases de datos de secuencias locales, esto resulta útil a la hora de reducir la carga de trabajo del usuario. Lazypipe implementa una arquitectura paso a paso flexible que permite la reejecución de pasos individuales o partes del análisis. Esta arquitectura aborda el mayor riesgo de falla de ejecución que es inherente al análisis de grandes bibliotecas NGS. Lazypipe admite formatos de datos que pueden ser utilizados tanto por investigadores humanos como por herramientas automatizadas.[61]

Después del proceso de estos genes se pueden aplicar diferentes estrategias para la anotación funcional de los ORF según el genoma o los genomas en cuestión. Para fagos comunes u homólogos cercanos en bases de datos públicas, las búsquedas de consultas en bases de datos de secuencias son utilizando BLAST o DIAMOND, pero gracias a la alta tasa de mutación de los fagos, es común que no se obtengan resultados significativos después de una búsqueda de similitud de secuencia, por lo que es mejor usar métodos de detección de homólogos remotos basados en modelos ocultos de Markov (HMM), que aprovechan el uso de perfiles de secuencia y la información de conservación de cada residuo estas bases de datos HMM de perfil múltiple se pueden usar junto con HMMER o HH-suite.

Aunque no todos los perfiles ofrecidos tienen una anotación funcional, la identificación de aciertos significativos contra esos modelos implica la identificación de genes que están presentes y conservados en otros genomas virales.

Además, estas bases de datos de perfil HMM también se pueden utilizar para explorar la historia evolutiva de los virus y sus proteínas. Tomando en cuenta la identificación de genes y sus funciones, la anotación de un genoma del fago abarca la anotación de ARN, repeticiones en tándem y retro elementos generadores de diversidad.[62]

A continuación, en la tabla 6 se mencionarán algunas bases de datos y sus respectivas descripciones.

Tabla 6. Descripción de las bases de datos de proteínas utilizadas para la anotación funcional de los ORF predichos

Base de datos	Descripción	Type[a]
RefSeq viral	Base de datos NCBI seleccionada de genomas virales, genes y proteínas. Actualizado periódicamente.	Secuencias
UniProtKB	Colección seleccionada de proteínas y proteomas de todos los dominios de la vida, derivados de envíos directos o predicciones del Archivo Europeo de Nucleótidos (ENA), GenBank y el banco de datos de ADN de Japón (DDBJ). La búsqueda BLAST está disponible en línea. Actualizado periódicamente.	Secuencias
Viral Eggnog	Agrupaciones de proteínas virales ortólogas derivadas de agrupación no supervisada basada en gráficos. Última actualización en 2016.	MMM
ViPhOG	Base de datos de grupos de dominios proteicos virales y fagos ortólogos generados a través del algoritmo CogSoft. Última actualización en 2021.	HMM y MSA
pVOG	Familias de genes de fagos derivadas del agrupamiento ortólogo de proteínas de fagos de genomas de fagos completos. Última actualización en 2016.	HMM y MSA
NCBI_CD	Colección de dominios bacterianos conservados, compilados a partir de seis bases de datos diferentes. La búsqueda web está disponible (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Última actualización en 2020.	HMM y PSSM
SCOP	Base de datos de proteínas conservadas estructural y evolutivamente, organizadas en una clasificación jerárquica de familias y superfamilias. También están disponibles dominios conservados basados en diferentes grados de identidad de secuencia. Última actualización en 2021.	Secuencias
VOGdb	Base de datos de grupos de proteínas virales ortólogas, derivadas del uso combinado del algoritmo CogSoft y la suite HH en genomas de fagos y profagos RefSeq. La base de datos proporciona proteínas específicas de virus (para la detección de secuencias virales en metagenomas) y paneles de proteínas virales esenciales. Actualizado periódicamente.	HMM y secuencias
VPF	Base de datos derivada del análisis Earth Virome de Paez-Espino et al., que consta de grupos de proteínas ortólogas virales. Última actualización en 2016.	HMM
FROG	Grupos homólogos remotos de proteínas de genomas completos de virus que infectan bacterias o arqueas. Última actualización en 2021.	HMM y MSA

Smits, S. L.; Bodewes, R.; Ruiz-Gonzalez, A.; Baumgärtner, W.; Koopmans, M. P.; Osterhaus, A. D. M. E.; Schürch, A. C. Assembly of Viral Genomes from Metagenomes. *Front. Microbiol.* **2014**, *5* (DEC), 1–10. https://doi.org/10.3389/fmicb.2014.00714.

Clasificación taxonómica

El proceso de clasificación taxonómica de muestras metagenómicas virales es más desafiante que el de los organismos celulares. La ascendencia común compartida de estos últimos permite la existencia de genes marcadores universales, en particular algunos genes, que pueden proporcionar una representación razonable de su origen evolutivo y divergencia, es así que este enfoque no es completamente aplicable a los virus, ya que carecen de un conjunto equivalente de genes

universalmente conservados, es así que una opción viable sería enfocarnos por clases de virus y sobre los cuales construir una filogenia, esta clasificación viral está limitada por dos factores. En primer lugar, las bases de datos de genomas virales actuales no reflejan la diversidad real de estos elementos en la biosfera y en segundo lugar, la taxonomía viral, tal como la define el Comité Internacional de Taxonomía de Virus (ICTV), se encuentra actualmente en un estado de cambio, con cambios que se están realizando y/o proponiendo, entre otros, al número de rangos taxonómicos a considerar, los criterios para definir cada rango específico y la validez de los caldos virales tradicionales, no basados en criterios moleculares. Sin embargo, en los últimos años han surgido varias herramientas para clasificar virus ensamblados a partir de metagenomas la mayoría de ellos basados en la similitud de secuencia mediante el uso de BLAST o HMM.[62]

Los enfoques basados en BLAST:

Para la asignación taxonómica de genomas virales usan BLAST para identificar similitudes entre los genomas de entrada y los genomas de referencia con taxonomía conocida como se ve en la tabla 6. VIRIDIC alinea todos los genomas en un conjunto de datos proporcionado por el usuario a través de BLASTn y luego procede a utilizar un algoritmo de agrupación jerárquica para agrupar los genomas virales en función de sus similitudes de alineación. VICTOR deriva distancias intergenómicas a partir de comparaciones por pares de secuencias de nucleótidos o aminoácidos de genomas virales completos o parciales. VipTree utiliza tBLASTx para determinar similitudes por pares de secuencias proteómicas putativas de un genoma de consulta frente a genomas virales de referencia incluidos en la herramienta y genera un árbol filogenético basado en dichas similitudes. La utilidad de estas tres herramientas está limitada por el grado de completitud de la consulta de contigs virales, el conocimiento previo que pueda tener el autor sobre su taxonomía y la base de datos empleada.[38]

Enfoques basados en Modelos de Markov

Los enfoques basados en Markov para la asignación taxonómica de genomas virales utilizan HMM y agrupación de Markov para identificar similitudes entre los genomas de entrada y los genomas de referencia con taxonomía conocida.

- VPF-Class ejecuta una búsqueda de HMM en tres conjuntos de datos diferentes para determinar la clasificación viral y la predicción del huésped.
- GRAViTy obtiene HMM de perfiles proteicos a partir de agrupaciones de proteínas ortólogas basadas en BLASTp y las utiliza para escanear genomas virales completos para obtener información sobre su contenido y orientación génicos.
- VIRify emplea información de presencia/ausencia de ViPhOG para la clasificación de contigs virales metagenómicos.
- Classiphage aprovecha un enfoque combinado de construcción de HMM, agrupamiento de cadenas de Markov y refinamiento de BLASTp para identificar HMM únicos para cada grupo de interés.
- vConTACT construye grupos de proteínas de fagos mediante el agrupamiento de Markov y genera similitudes por pares entre genomas.[63]

Tabla 7. Comparación de las herramientas para la clasificación taxonómica de datos metagenómicos de fagos

Herramienta	Acercarse	Accesibilidad	Actualizado recientemente
VIRIDIC	BLAST seguido de agrupación jerárquica	Versión independiente	Sí
VÍCTOR	Distancias intergenómicas derivadas de explosiones	Servicio web en línea	No
VipTree	tBLASTx	Servicio web en línea	Sí
Método Dougan y Quake	distancias tBLASTx y 4-mer	Debe ser implementado por el usuario	N / A
Clase VPF	HMM contra diferentes bases de datos	Versión independiente	Sí
Gravity	Presencia/ausencia y sintenia de grupos ortólogos determinados a través de HMM	Versión independiente	Sí
VIRificar	HMM de ViPhOG	Servicio web en línea	Sí
Clasifago	HMM refinados por BLASTp	Base de datos HMM disponible para descargar, las distancias deben implementarse automáticamente	Sí
vContact2	Distancias derivadas de grupos de proteínas de fagos basados en Markov	Versión independiente	Sí

Andrade-martínez, J. S.; Camelo, C.; Chica, A.; Forero-junco, L.; López-leal, G.; Moreno-gallego, J. L.; Rangel-pineros, G. Computational Tools for the Analysis of Uncultivated Phage Genomes.

Capítulo 5. Análisis filogenético y evolución molecular

Algo clave dentro del estudio de datos en las secuencias moleculares de virus es la filogenética, que es el estudio de la relación evolutiva en varios grupos de organismos o entidades biológica. En el contexto del estudio de virus, la filogenética es utilizada para analizar y comparar secuencias moleculares, como el ADN o el ARN viral. Al analizar las secuencias genéticas de diferentes cepas o variantes de un virus, se pueden inferir su relación evolutiva y determinar cómo han evolucionado y divergido a lo largo del tiempo.[64]

El origen y la evolución de los virus siguen siendo difíciles de explicar debido a la falta de consideración de la composición completa de los proteomas virales. Quizás los problemas más desafiantes que plagan los estudios evolutivos profundos de los virus son la rápida evolución y las altas tasas de mutación de la mayoría de los genes virales, en específico los virus de ARN. Esto dificulta la unificación de las familias virales, especialmente mediante el análisis filogenético basado en secuencias, además se espera que el número de familias virales sin orden aumente continuamente, especialmente con el descubrimiento de nuevos virus de ambientes atípicos y porque los genes de muchas familias virales no exhiben similitudes de secuencia significativas, esto que las proteínas homólogas a menudo divergen más allá del reconocimiento a nivel de secuencia después de que ha pasado un tiempo evolutivo relativamente largo.[65]

5.1. Métodos de alineación de secuencias

El método de alineación de secuencias para análisis filogenético es un paso muy útil, pero a la vez costoso hablando computacionalmente debido al creciente aumento de los datos de secuencia de ADN, estando estos datos muy fragmentados, anteriormente estas alineaciones producían algoritmos automatizados que después son verificados, sin embargo esto ya es ineficaz en la filogenética moderna y da como resultado alineaciones que no los reproducibles, por lo que se ha empleado métodos para automatizar por completo las alineaciones de grandes conjuntos de datos, se pueden resumir que en este método hay tres categorías, un enfoque automatizado usando software tradicional, un enfoque automatizado que incluya edición manual y visual y enfoques totalmente automatizados, para comprender cómo afectan estos

métodos de resultados filogenéticos, se comparan las topologías de árboles resultantes en función de estos diferentes métodos de alineación usando múltiples métricas.[66]

Existen dos tipos de alineación de secuencias: el alineamiento de secuencias en pares (PSA) y el alineamiento de secuencias múltiples (MSA). El MSA es más ventajoso que el PSA, esto dependiendo del contexto biológico, ya que considera varios miembros de una familia de secuencias y, por lo tanto, proporciona más información biológica. El MSA también es un requisito previo para análisis genómicos comparativos, que permiten la identificación y cuantificación de regiones conservadas o motivos funcionales en una familia de secuencias completa, la estimación de la divergencia evolutiva entre secuencias e incluso el perfilado de secuencias ancestrales.[67]

El método de alineación de secuencias (SA) se puede dividir en dos categorías: alineamiento global y alineamiento local. El alineamiento global se lleva a cabo cuando se evalúa la similitud en toda la extensión de las secuencias. Varios métodos de MSA logran el alineamiento global, pero surgen dificultades cuando las secuencias solo son homólogas en regiones locales, donde existe un bloque claro de alineamiento sin brechas que es común a todas las secuencias, o cuando hay presencia de dominios reorganizados entre las secuencias relacionadas. En tales casos, se realiza el alineamiento local para identificar las regiones similares específicas en las secuencias. Cuando hay una gran diferencia en las longitudes de las secuencias a comparar, generalmente se prefiere realizar el alineamiento local.[67]

Alineamientos pareados

El alineamiento de secuencias en pares es la base sobre la cual se construye el alineamiento de múltiples secuencias, y se divide principalmente en dos tipos: alineamiento local y alineamiento global. El alineamiento local se centra en encontrar y alinear regiones similares en las secuencias, mientras que el alineamiento global se ocupa de realizar una alineación completa desde el principio hasta el final de las secuencias. Un algoritmo ampliamente utilizado para el alineamiento global es el algoritmo de Needleman-Wunsch, que se ha convertido en el algoritmo básico utilizado en diversos programas de software de alineamiento de múltiples secuencias.

Este algoritmo generalmente se compone de dos pasos: el primero implica calcular los estados de una matriz de programación dinámica, y el segundo implica realizar un seguimiento desde el estado final hasta el estado inicial de dicha matriz para obtener la solución de alineamiento.[68]

Múltiples secuencias

El cálculo de una alineación de múltiples secuencias (MSA) exacta es un problema computacionalmente complejo y, por lo tanto, el método exacto no se utiliza en la práctica para conjuntos de datos realistas debido a su dificultad. En su lugar, la mayoría de los procedimientos de MSA son aproximaciones o heurísticas que ofrecen soluciones de alineación factibles dentro de un tiempo limitado. Dado que las heurísticas existentes no siempre proporcionan la mejor solución y debido al crecimiento constante del tamaño de las bases de datos, actualmente se investiga el desarrollo de nuevos métodos de MSA con un alto rendimiento para obtener alineaciones de secuencias precisas. Existen tres categorías principales de enfoques utilizados en MSA: los enfoques exactos, progresivos y los iterativos.

- Exactos

Es el método más directo para producir alineamientos múltiples de secuencias utiliza la técnica de programación dinámica para identificar la solución de alineamiento globalmente óptima. Para secuencias de nucleótidos puede usarse una matriz de sustitución, pero dado que solo hay cuatro caracteres estándar posibles por secuencia, y que los nucleótidos individuales no difieren mucho en su probabilidad de sustitución

- Progresivos

Construyen un alineamiento múltiple final realizando una serie de alineamientos de pares sobre secuencias sucesivamente menos emparentadas tales métodos comienzan alineando en primer lugar las dos secuencias más cercanamente relacionadas, para seguir alineando sucesivamente la siguiente secuencia del conjunto problema más emparentada con el alineamiento producido en el paso previo, es así que construyen automáticamente tanto un árbol filogenético como un alineamiento

- Iterativos

Los métodos son un conjunto de métodos para producir alineamientos múltiples de secuencias que reducen los errores inherentes en los métodos progresivos pero este enfoque mejora la eficiencia a costa de la precisión. Los enfoques exactos tienden a proporcionar alineaciones de alta calidad que se acercan mucho a lo óptimo.

5.2. Construcción de árboles filogenéticos

Recientemente, ha sido una dificultad poder reconstruir la historia evolutiva de los virus, en promedio un virus se construye usando fragmentos de códigos de uno o más virus que son sus antepasados, la construcción de estos árboles filogenéticos se puede clasificar en tres categorías principales, métodos de parsimonia, distancia, máxima verosimilitud y bayesianos. En el método de distancia, si dos organismos que comparten son cien puntos en común o son más similares entre sí que dos organismos cuyo último punto en común o fue más lejano. Por esta razón, debería ser posible inferir relaciones evolutivas a partir de similitudes encontradas entre organismos. Recordemos que las distancias representan la disimilitud entre cada par de organismos que representan las taxas. A partir de las distancias se genera una matriz de distancias, la que se utiliza para generar un árbol filogenético.[69]

- Parsimonia

El método de parsimonia mediante este método se obtienen árboles que ordenan las ramas de modo tal que se minimiza el número de mutaciones que deben haber ocurrido en el tiempo.

- Verosimilitud

En el método de máxima verosimilitud se calcula la probabilidad de que nuestros datos hayan generado distintos árboles posibles y se devuelve el árbol que presenta una máxima probabilidad

- Bayesianos

Que calcula una probabilidad posterior para cada árbol posible dado un modelo de evolución y unas observaciones. Es decir, dadas unas observaciones, la inferencia

bayesiana actualiza las probabilidades de que los árboles sean correctos, y una herramienta que se ha vuelto cada vez más popular en los últimos años es BEAST, que utiliza árboles filogenéticos con tiempo medido para análisis evolutivos bayesianos, como la genética de poblaciones basada en el coalescente, la filodinámica y la filogeografía.[70]

5.3. Métodos para el análisis de recombinación

La recombinación genética es un proceso evolutivo importante que genera gran parte de la diversidad genética sobre la que actúa la selección natural. Los patrones de recombinación que son evidentes dentro de los genomas de dichos virus pueden revelar mucho sobre su biología y evolución. Los patrones no aleatorios de intercambio de secuencias entre individuos dentro de una especie pueden proporcionar evidencia directa de subdivisiones de población impuestas por el rango de huéspedes o geográficas que evitan que ciertos individuos se recombinen, es así que los patrones de intercambio de secuencias entre virus en diferentes especies pueden revelar vínculos ecológicos que de otro modo serían indetectables entre algunas especies y barreras entre otras.

Debido a la gran variedad, es necesario tener estrategias para generar diversidad genética, así que sufren los virus cambios genéticos como lo son la mutación puntual y la recombinación, la recombinación ocurre cuando mínimo dos genomas virales coinfectan la misma célula del hospedero y a la vez intercambian segmentos genéticos, hay varios tipos de recombinación como lo son la recombinación homologa que ocurre en el mismo sitio en ambas cadenas parentales, no la homologa, originando frecuentemente estructuras aberrantes y una no muy común la recombinación por reordenamiento que ocurre en el virus con genomas segmentados, que pueden intercambiar segmentos completos del genoma, esta recombinación es muy útil, ya que es un fenómeno generalizado en los virus y puede tener un gran impacto en su evolución.

Dada la frecuencia, relevancia e impacto de la recombinación en la evolución viral, no sorprende que se hayan desarrollado muchos enfoques bioinformáticos para detectar y estimar con precisión su ocurrencia en los genomas virales. Algunas de estas herramientas ya aprovechan los datos genómicos generados mediante la

secuenciación de alto rendimiento. No obstante, las nuevas oportunidades que ofrecen estas tecnologías también vienen con algunos desafíos analíticos serios que deben ser considerados.[71]

El no estimar la recombinación puede llevar a la mala cuantificación de las presiones de selección y las estimaciones filogenéticas, es esencial realizar un análisis de detección de recombinación en los análisis filogenéticos. Los métodos para detectar la recombinación se pueden dividir en cuatro categorías:

- Los métodos de distancia utilizan las diferencias genéticas entre secuencias en diferentes posiciones a lo largo del genoma para identificar la presencia de recombinación.
- Los métodos filogenéticos exploran inconsistencias entre las topologías de árboles de diferentes partes del genoma.
- Los métodos de compatibilidad son enfoques filogenéticos que prueban, sitio por sitio, si cada sitio es compatible con el mismo árbol.
- Los métodos de distribución de sustituciones evalúan el ajuste a una distribución estadística esperada o la agrupación significativa de sustituciones.

Una vez que se ha tenido en cuenta la recombinación, es posible que se desee reconstruir la dinámica evolutiva y epidemiológica de la parte no recombinante del genoma viral. La reconstrucción filogenética puede ser basada en distancia (por ejemplo, vecino más cercano) o basada en caracteres (por ejemplo, parsimonia máxima, máxima verosimilitud o inferencia bayesiana).

En los genomas virales las distribuciones de puntos de ruptura de recombinación que son evidentes y se pueden revelar detalles de los procesos mecánicos y bioquímicos que subyacen a la recombinación, las fuerzas selectivas que limitan la supervivencia y proliferación de los recombinantes, de aquí viene la importancia de la recombinación, otra razón importante para analizar los patrones de recombinación en los genomas de virus es minimizar el impacto perturbador que la recombinación puede tener en otros análisis de evolución molecular basados en la filogenia.

Una herramienta basada en métodos filogenéticos es RDP4 es la última versión del programa de detección de recombinación (RDP), un programa informático de Windows que implementa una amplia gama de métodos para detectar y visualizar la recombinación y eliminar la evidencia de recombinación de las alineaciones de secuencias del genoma del virus. RDP4 es capaz de analizar el doble de secuencias que son hasta tres veces más largas que las que podían analizar las versiones anteriores del programa. La característica clave de RDP4 que lo diferencia de otras herramientas de detección de recombinación es su flexibilidad. Se puede ejecutar en modo totalmente automatizado desde la interfaz de línea de comandos o con una interfaz de usuario rica en gráficos que permite la exploración detallada tanto de los eventos de recombinación individuales como de los patrones generales de recombinación.[72]

Por otro lado, otro programa para la detección de recombinación GARD es un algoritmo genético para la detección de recombinación implementa un enfoque de algoritmo genético para detectar alineaciones en busca de evidencia de incongruencia filogenética, que se interpreta como un sello distintivo de recombinación, conversión de genes o procesos similares. La salida de GARD es una alineación dividida en formato NEXUS, con árboles específicos de partición, que representan fragmentos supuestamente no recombinantes.[72]

5.4. Métodos para el análisis de selección natural

La selección natural actúa sobre los organismos celulares asegurando que los genes responsables den un fenotipo ventajoso. Esto es posible porque las células reproductivas de estos organismos son casi siempre haploides, separando el gen beneficioso de su alelo rival en cada generación.

En el caso de los virus, realizan copias del genoma viral por medio de la replicación, creando así el entorno copias del genoma viral, donde algunas de ellas experimentan cuello de botella, lo que permite que la selección natural recompense las mutaciones beneficiosas y, purgue los errores letales.

La selección actúa sobre las diferencias fenotípicas, esto simplemente significa que, dentro de una población de la misma especie que vive en condiciones casi idénticas,

algunos individuos pueden manifestar ciertas características físicas o de comportamiento, fenotipos, que les permiten superar a otros, esta regla de la selección natural es que la diferencia fenotípica bajo la selección debe ser heredable, lo que significa que debe deberse a diferencias en genes específicos, con frecuencia en forma de variantes alélicas del mismo gen ancestral.

El aumento de la frecuencia del gen beneficioso en la población, a su vez, permite que más individuos expresan un fenotipo ventajoso, perpetuando el ciclo virtuoso de la selección natural. La selección purificadora no puede ocurrir a menos que la mutación letal afecte a los alelos funcionales.

De hecho, la mayoría de las enfermedades hereditarias causadas por genes mutantes con pérdida de función, son recesivas y manifiestan sus fenotipos fatales solo cuando los genes mutados son homocigotos, estando completamente separados de sus alelos funcionales.[73]Por lo tanto, la selección natural es un proceso evolutivo que se caracteriza por la supervivencia diferencial de individuos con diferentes genotipos en una población.[74]

Las mutaciones que ocurren en el genoma pueden o no alterar el fenotipo de un organismo, lo que a su vez puede afectar su adaptación. La mayoría de las nuevas mutaciones que surgen en una población disminuyen la adaptabilidad de los individuos que las portan, y se conocen como mutaciones deletéreas.

Estas mutaciones son seleccionadas en contra y eventualmente eliminadas de la población. Este tipo de selección se conoce como selección negativa o purificadora. Sin embargo, en ocasiones, una mutación puede aumentar la adaptabilidad de los individuos que la poseen.

A esta mutación se le considera ventajosa, y la fuerza que impulsa su incremento en la frecuencia en la población se conoce como selección positiva o diversificadora.[75]

Los enfoques más ampliamente empleados para detectar la existencia de selección se basan en la observación de la relación dN/dS o ω, la cual compara las tasas de sustituciones no sinónimas por sitio no sinónimo (dN) con las tasas de sustituciones sinónimas por sitio sinónimo (dS). Estos métodos parten del supuesto de que las

mutaciones sinónimas son neutrales, es decir, no afectan la aptitud del individuo que las porta.

Por lo tanto, los valores obtenidos de la relación dN/dS se interpretan de la siguiente manera:

dN/dS = 1 indica neutralidad.

dN/dS < 1 sugiere selección negativa.

dN/dS > 1 indica selección positiva.

Los métodos modernos de dN/dS utilizan la información filogenética y la reconstrucción de caracteres ancestrales, además de modelos estadísticos, para determinar el número y tipo de sustituciones que han ocurrido en los codones. También tienen la capacidad de calcular el dN/dS de forma individual para cada codón, así como para cualquier conjunto de codones presentes en un alineamiento, lo que permite localizar con precisión la adaptación molecular de un gen.[76]

Identificar las presiones de selección que han moldeado la evolución molecular de un virus. Los métodos disponibles para esto se pueden dividir en tres clases:

- Métodos de conteo que enumeran el número de sustituciones no sinónimas y sinónimas a lo largo de la filogenia.
- Modelos de efectos aleatorios que asumen una distribución de tasas en diferentes sitios y deducen las tasas de sitios individuales según la distribución.
- Modelos de efectos fijos que estiman la tasa de sustituciones no sinónimas a sinónimas con base en cada sitio.

Una consecuencia de las fuerzas evolutivas ha ocasionado una diversidad genética existente, siendo esta una problemática para el análisis comparativo en secuencias modernas, estos avances en la generación de secuencias y la mayor sofisticación estadística de los métodos relevantes ahora permiten a los investigadores extraer cada vez más señales evolutivas de los datos, aunque a un mayor costo computacional. Para ellos existen diferentes herramientas como Hyphy o Daratamoney y PAML.[77]

Hyphy es la versión stand alone de Datamonkey 2.0, una versión completamente rediseñada del servidor web de Datamonkey para analizar formas evolutivas en datos de secuencia. Datamonkey 2.0 proporciona una colección cuidadosamente seleccionada de métodos para detectar huellas de selección natural.[77]

Dentro de Datamonkey 2.0 incluye pruebas estadísticas para caracterizar las fuerzas evolutivas que han actuado en la alineación de secuencias, que son métodos diseñados para análisis exploratorios y pruebas, estos son:[78]

- SLAC

El conteo de antepasados de probabilidad única o (Single Likelihood Ancestor Countin en inglés) SLAC es un método basado en el conteo de sustitución para identificar sitios que pueden haber experimentado una selección diversificadora o purificadora generalizada. Es capaz de manejar conjuntos de datos más grandes, pero generalmente tiene el poder estadístico más bajo de todos los métodos específicos del sitio.

- FEL

La probabilidad de efectos fijos, o FEL es un método de máxima verosimilitud utilizado para identificar sitios que pueden haber experimentado una selección diversificadora o purificadora generalizada al probar individualmente ($dN/dS \neq 1$) en cada sitio de la alineación.

- BUSTED

La prueba estadística sin restricciones de sitio de rama para la diversificación episódica, o BUSTED es una prueba de razón de verosimilitud para la evidencia de selección diversificadora que afecta algunos sitios no especificados en la alineación a lo largo de algunas ramas no especificadas en el árbol.

Es más adecuado para la detección de alineaciones relativamente pequeñas donde los métodos centrados en el sitio o en la rama tienen poco poder estadístico para detectar eventos selectivos locales. Al agregar señales en sitios y ramas del árbol, BUSTED puede lograr una mayor potencia en pequeños conjuntos de datos.

- FUBAR

  La aproximación bayesiana rápida sin restricciones, o FUBAR es un método diseñado para identificar sitios que pueden haber experimentado una selección generalizada de diversificación o purificación. Utiliza una aproximación estadística basada en principios para limitar el número de costosas evaluaciones de probabilidad y es adecuado para grandes conjuntos de datos.

- MEME

El Modelo de Evolución de Efectos Mixtos, o MEME incluye una prueba de razón de verosimilitud para detectar sitios individuales sujetos a una selección diversificadora episódica. A diferencia de SLAC, FEL y FUBAR, MEME puede identificar sitios donde solo algunas de las sucursales han experimentado presión selectiva. Con un conjunto de datos lo suficientemente grande, MEME proporciona la mayor potencia en Datamonkey 2.0 para detectar la selección a nivel de sitio

- ABSREL

El método adaptativo Branch-Site Random Effects Likelihood, o aBSREL evalúa la evidencia de selección positiva que afecta a ramas individuales, que se puede especificar *a priori* o examinar exhaustivamente. El método también determina qué ramas admiten modelos evolutivos más complejos y, por lo tanto, se puede utilizar para una estimación más precisa de las longitudes de las ramas en secuencias de codificación heterotachous para la datación del reloj molecular

- RELAX

Este es un método especializado que prueba formalmente si la selección se ha relajado o intensificado en una colección de ramas especificadas *a priori* en el árbol en relación con otras. RELAX también es más útil en general para comparar regímenes selectivos en diferentes partes del árbol

Por otro lado la teoría neutral de la evolución molecular, donde dice que la mayor parte de la variación observada dentro y entre especies no se debe a la selección natural, sino más bien a la fijación aleatoria de mutaciones con poca importancia para el medio, entonces estas mutaciones suelen ser ventajosas, pero son raras a nivel

molecular, estas mutaciones ventajosas en genes y genomas hacen posible dar forma a la morfología, el comportamiento y la fisiología de las especies, o de las divergencias de las especies y las innovaciones evolutivas.

Detectar la adaptación molecular nos permite así lograr una mejor comprensión del proceso evolutivo. Los genes que codifican proteínas, se puede diferenciar las sustituciones sinónimas o silenciosas, que son sustituciones de nucleótidos que no cambian el aminoácido codificado de las sustituciones no sinónimas o de reemplazo, estas son sustituciones que sí cambian el aminoácido. Debido a que la selección natural opera principalmente a nivel de proteína, las mutaciones sinónimas y no sinónimas están bajo presiones selectivas muy diferentes y fijadas a ritmos muy diferentes.

Es así que con la tasa de sinónimos actuando como punto de referencia, se puede decir si la fijación de mutaciones no sinónimas en la población se acelera o se desacelera por la selección natural que actúa sobre la proteína. Por lo que la comparación de las tasas de sustitución sinónimas y no sinónimas puede revelar la dirección y la fuerza de la selección natural que actúa sobre la proteína, un gen con una tasa acelerada de sustitución no sinónima, indicada por la relación de tasa no sinónima/sinónima dN /dS > 1, se dice que está bajo selección positiva, debido a que son datos de secuencias codificantes de nucleótidos.

Este tipo de prueba es particularmente eficaz para detectar selección diversificada o selección equilibrada, ya que utiliza sustituciones no sinónimas excesivas como evidencia de que la selección natural ha ayudado a la fijación de mutaciones no sinónimas. Las pruebas basadas en dN / dS pueden ser menos efectivas cuando se aplican a datos de la misma especie debido a la falta de divergencias en la secuencia y a las complicaciones en la interpretación de la relación dN / dS. Es por ello que CODEML en el paquete PAML se ha utilizado ampliamente para analizar secuencias de genes que codifican proteínas para estimar las tasas sinónimas y no sinónimas (dS y dN) y para detectar la selección darwiniana positiva que impulsa la evolución en las secuencias codificantes.[78]

5.5. Métodos para análisis de coevolución

Una coevolución se puede referir a patrones compensatorios mutacionales en donde se preserva la función de la proteína, las glicoproteínas que tiene la envoltura viral, median la entrada de los virus envueltos en sus células huésped, estas están formadas por señales de coevolución que le dan a los virus cierta plasticidad para evadir anticuerpos neutralizantes sin alterar mecanismos de entrada viral, es por ello que con un análisis de coevolución se puede predecir importantes características estructurales y reordenamientos de complejos de proteínas virales mediante análisis computacionales.[79]

La detección de residuos coevolutivos en una proteína mediante el análisis comparativo de secuencias génicas homólogas es una importante fuente de evidencia para la caracterización funcional y/o estructural de las proteínas, es así que el análisis comparativo de secuencias de nucleótidos no codificantes puede revelar una estructura secundaria, pero al no abordar la naturaleza evolutiva de la variación de la secuencia, tales métodos son susceptibles a asociaciones espurias entre sitios debido a la identidad por descendencia y que las pruebas de asociación por pares no pueden capturar interacciones de orden superior y no proporcionan un medio para compilar el panorama general de una lista de pares significativos.

La herramienta Spidermonkey proporciona una interfaz web fácil de usar para un marco para detectar sitios coevolutivos a partir de secuencias de nucleótidos o proteínas codificantes y no codificantes, que combina técnicas filogenéticas y de aprendizaje automático para abordar estos problemas, la historia de los eventos de sustitución se infiere a partir de una alineación utilizando métodos filogenéticos estándar.

Si un árbol no se carga con la alineación, entonces se estima utilizando el método de unión de vecinos, los conjuntos replicados de secuencias ancestrales se pueden volver a muestrear a partir de la distribución de probabilidad posterior y analizarse en paralelo. Para los datos de codones, solo las sustituciones no sinónimas se conservan para análisis posteriores. Los sitios invariantes se excluyen automáticamente en todos los casos. Los patrones correlacionados de sustituciones en el árbol implican una coevolución entre sitios. La distribución conjunta de sustituciones en el árbol se

codifica como una matriz de estado binaria, en la que cada fila corresponde a una rama única y cada columna a un sitio en la alineación, y se analiza mediante modelos gráficos bayesianos (BGM). Un BGM es una representación compacta de una distribución de probabilidad conjunta en la que cada nodo representa una variable aleatoria distinta.

Es por ello que Spidermonkey es un componente del conjunto de herramientas filogenéticas de Datamonkey que proporciona métodos para detectar sitios coevolutivos a partir de una alineación múltiple de secuencias homólogas de nucleótidos o aminoácidos. Reconstruye el historial de sustitución de la alineación mediante métodos filogenéticos basados en la máxima verosimilitud y luego analiza la distribución conjunta de los eventos de sustitución utilizando modelos gráficos bayesianos para identificar asociaciones significativas entre sitios.[78]

Capítulo 6. Herramientas especializadas para la Búsqueda de hospederos virales

Predecir la asociación virus-huésped es esencial para comprender cómo interactúan los virus con las especies huésped y descubrir nuevas terapias para enfermedades virales en humanos y animales. Las herramientas computacionales permiten un descubrimiento rutinario de virus previamente desconocidos en muestras metagenómicas de una amplia gama de entornos, herramientas computacionales para predecir huéspedes procarióticos a partir de secuencias de virus derivadas del metagenoma en función de las señales moleculares de la coevolución virus-huésped, incluida la homología de secuencias, la similitud de la composición de secuencias entre los virus y sus huéspedes.[80]

Se han desarrollado varias herramientas computacionales para predecir el hospedero de un virus mediante el análisis de su secuencia de ADN o ARN, las coincidencias con las secuencias espaciadores del CRISPR (en inglés: Clustered Regularly Interspaced Short Palindromic Repeats). Estos métodos se pueden dividir en tres enfoques generales: aprendizaje supervisado, modelos probabilísticos y clasificaciones de similitud. Todos estos enfoques requieren características con las que se pueda clasificar la secuencia de entrada. Las características utilizadas para la clasificación son principalmente *k-mer* basadas en varios tamaños de *k* entre 1 y 8.

En el caso de modelos probabilísticos y clasificaciones de similitud, no solo se deben analizar los genomas virales, sino también los genomas del hospedero.[81]

Los métodos computacionales para predecir el o los hospederos de los fagos en función de su secuencia genómica actualmente son objeto de investigación activa, las herramientas existentes de predicción del hospedero aprovechan varios niveles y patrones de similitud de secuencia entre los genomas del fago y del huésped o utilizan un enfoque comparando el fago consultado con una base de datos de virus con virus conocidos.

En las herramientas basadas en el huésped, la similitud de secuencia entre los genomas del fago y del huésped puede detectarse a través de la alineación de secuencias, alternativamente, las herramientas basadas en el huésped pueden basarse en enfoques sin alineación. Por el contrario, las herramientas basadas en fagos no se basan en la similitud fago-huésped, sino que extraen información de una base de datos de fagos y arqueovirus de referencia con hosts conocidos, se han desarrollado varias herramientas automatizadas que aprovechan el aprendizaje automático para obtener una predicción de host integrada.

La identificación del huésped viral es esencial para la caracterización de los fagos, ya que dependen del huésped para su replicación. Actualmente, el método más común utilizado para determinar el huésped de un fago es a través de cultivos, pero esto puede ser ineficiente, lento y costoso, hay varios enfoques para predecir las relaciones fago-huésped; en su mayor parte, se basan en perfiles de abundancia, homología genética, CRISPR, coincidencias exactas o perfiles de oligonucleótidos estos se dividen en dos grupos principales en función de su uso de alineación de secuencias: métodos basados en alineación o sin alineación.

Los métodos basados en alineación se basan en búsquedas de similitud de secuencia entre un virus de consulta y un genoma huésped, ya que los virus y los huéspedes pueden compartir genes y secuencias de nucleótidos cortas. Los métodos sin alineación predicen el huésped de un virus en función de los *k-meros*, concurrentes de los fagos con huéspedes conocidos o la similitud de las firmas de secuencia entre los virus y sus huéspedes.[82]

La agrupación de los datos virales se ha realizado utilizando los enfoques tradicionales basados en secuencias o enfoques basados en redes. Uno de los programas para la predicción de hospederos es WIsH predice huéspedes procarióticos de fagos a partir de secuencias contig de fagos. WIsH opera mediante la comparación de secuencias genómicas de fagos con las secuencias genómicas de hospedadores procariotas conocidos. Utiliza un método estadístico para determinar la probabilidad de que un fago tenga su origen en un hospedador procariota específico. WIsH ha demostrado tener una gran precisión en la predicción de los hospedadores de fagos. Al predecir el género del hospedador dentro de 20 géneros para contigs de fagos de 3 kbp de longitud, su precisión promedio alcanza el 63%. Esta precisión es significativamente mayor que la de otras herramientas utilizadas para predecir los hospedadores de fagos.[46]

La herramienta VirHostMatcher predice hosts bajo el supuesto de que los genomas del virus y del huésped a menudo tienen frecuencias de oligonucleótidos similares. Además, se han desarrollado un clasificador para distinguir entre fagos y virus que infectan a células eucariotas. Una de las estrategias desarrolladas para la detección viral y su posterior clasificación es considerar las proteínas virales como blanco. Las familias de proteínas virales (VPF) se han utilizado ampliamente en la identificación de nuevas secuencias virales en grandes conjuntos de datos.[50]

La herramienta VPF-Class, es una herramienta para clasificar genomas virales con respecto a la taxonomía y la predicción del huésped en múltiples niveles taxonómicos. Una de las ventajas de nuestra herramienta es proporcionar una asignación taxonómica, así como una predicción del huésped de cada genoma viral en lugar de una agrupación de datos virales.[83]

La herramienta HostPhinder que predice el huésped bacteriano de un fago en función de su secuencia genómica. Utiliza una base de datos de referencia de fagos con huéspedes conocidos y mide la similitud genética entre el fago consultado y los fagos de la base de datos. Cuanto más similares genéticamente sean dos fagos, más probable es que compartan el mismo huésped bacteriano.

HostPhinder fue desarrollado por un equipo de investigadores de la Universidad Técnica de Dinamarca. Se publicó por primera vez en 2016 y desde entonces se ha utilizado para predecir los huéspedes de una amplia variedad de fagos. HostPhinder se basa en una base de datos de referencia de fagos con hosts conocidos. La base de datos contiene las secuencias del genoma de más de 10.000 fagos y se actualiza periódicamente. HostPhinder utiliza esta base de datos para medir la similitud genética entre un fago de consulta y los fagos de la base de datos. Cuanto más similares genéticamente sean dos fagos, más probable es que compartan el mismo huésped bacteriano.

Se ha demostrado que HostPhinder es muy preciso en la predicción de hosts de fagos. En un estudio publicado en 2016, HostPhinder pudo predecir correctamente el huésped de un fago el 92 % de las veces. Esto convierte a HostPhinder en una herramienta valiosa para los investigadores que estudian fagos.[84]

VirusHostDB, es una base de datos para la búsqueda de hospederos, la base de datos contiene información sobre los anfitriones de más de 10.000 virus, incluidos: El tipo de virus (por ejemplo, bacteriófago, virus vegetal, virus animal), el organismo huésped (por ejemplo, bacterias, plantas, animales), el rango de hosts (es decir, el rango de hospederos que un virus puede infectar), La secuencia del genoma del virus y Las referencias bibliográficas para la asociación huésped-virus.

VirHostDB se puede utilizar para buscar virus que infecten a un organismo huésped en particular o para buscar hospederos que estén infectados por un tipo de virus en particular. La base de datos también se puede utilizar para visualizar el rango de host de un virus o para identificar posibles nuevos hosts para un virus. VirHostDB es un recurso valioso para los investigadores que estudian virus y sus anfitriones. Es una base de datos completa y actualizada que se puede utilizar para responder a una amplia variedad de preguntas sobre las interacciones virales con el huésped.

VirHostDB, es una base de datos completa que contiene información sobre los anfitriones de más de 10.000 virus. Se actualiza regularmente, por lo que puede estar seguro de que la información está actualizada. Es gratuito y de acceso abierto, por lo que cualquiera puede usarlo.

Los métodos sin alineación, de identificación de hospedero, usan los perfiles *de k*-mer, y utilizan la composición de nucleótidos para predecir el huésped de un genoma viral. Algunos virus adaptan su composición de oligonucleótidos a la del huésped que infectan, un proceso que puede estar impulsado por la adaptación del uso de codones a la maquinaria de traducción y al grupo de ARNt disponible en la célula huésped, el intercambio del material genético, la coevolución de secuencias reguladoras, y/o una evasión de los sistemas de defensa del huésped.

Así que es necesario identificar el genoma procarionte con la mayor similitud significativa con un genoma viral, las herramientas que explotan los perfiles de *k-mer* asumen que el genoma procariota es el huésped del virus en cuestión, existen enfoques dependientes de la alineación para evaluar la similitud entre los genomas viral y procariota. Estos métodos asumen que el intercambio de información genética entre genomas virales y procariotas es indicativo de asociaciones virus-huésped.

Los fragmentos genéticos específicos, aunque cortos, pueden ser informativos para este propósito, como los espaciadores CRISPR y los genes de ARNt, mientras que las coincidencias más largas, como los genes completos o los profagos integrados, también pueden proporcionar una indicación del vínculo virus-huésped, el contenido de genes de las secuencias virales se puede investigar en busca de genes marcadores específicos que sean indicativos del huésped, como los genes de fotosíntesis para cianófagos. Todos estos enfoques se han utilizado ampliamente en estudios metagenómicos virales para predecir huéspedes de virus no cultivados.[85]

EvoMIL es un método de aprendizaje profundo que predice la asociación virus-huésped a nivel de especie solo a partir de la secuencia viral. El método combina un modelo de lenguaje de proteínas grandes preentrenado y aprendizaje de instancias múltiples basado en la atención para permitir predicciones orientadas a proteínas. Cabe recalcar que EvoMIL también estima la importancia de las proteínas individuales en la predicción y las asigna a un paisaje incrustado de todas las proteínas virales, donde las proteínas con funciones similares se agrupan claramente.[86]

PhisDetector combina varios métodos basados en alineación como sin alineación, y utiliza un conjunto de enfoques de aprendizaje automático para evaluar la confianza

de cada posible par fago-huésped. VirHostMatcher-Net propone integrar tanto la señal de virus-virus como la de virus-huésped en una red modelada de virus-huésped, a partir de la cual se evalúan los posibles pares de virus-huésped mediante una regresión logística. Si bien ambas herramientas mostraron mejoras potenciales en comparación con los métodos individuales, ninguno de los puntos de referencia proporcionados sugirió que podrían alcanzar una tasa de descubrimiento falso (FDR) baja (<10 %) a nivel de género del huésped, incluso con los límites más estrictos.

iPHoP es un enfoque de aprendizaje automático de dos pasos que predice los hospedadores de virus derivados de metagenomas. El primer paso utiliza un enfoque basado en fagos para identificar genes específicos de un género de hospedador en particular. El segundo paso utiliza un enfoque basado en hospedadores para identificar genes compartidos por un grupo de géneros de hospedadores. El flujo de trabajo iPHoP fue evaluado en un conjunto de datos de 216,015 genomas de fagos derivados de metagenomas. Fue capaz de predecir el género del hospedador del 86% de los fagos con una tasa de falsos positivos del 10%,y puede utilizarse para identificar los hospedadores de nuevos fagos y para estudiar la ecología de los virus en el medio ambiente, además permite una predicción de género huésped de alta confianza (estimado <10 % FDR) para fagos en una amplia gama de ecosistemas. iPHoP es una herramienta poderosa para predecir los hospedadores de virus derivados de metagenomas.[87]

Capítulo 7. CRISPR-Cas y Virus

Las infecciones virales emergentes y reemergentes soy hoy un problema de salud global, un ejemplo reciente y claro, la pandemia por el nuevo coronavirus causante de la COVID-19, la identificación rápida, sensible, específica y desplegable en el campo de los virus actuales es fundamental para la vigilancia, el control y el tratamiento de enfermedades. Las deficiencias de los métodos actuales crean una necesidad inminente de desarrollar nuevas plataformas de biodetección. Sistemas CRISPR-Cas(Clustered Regularly Interspersed Short Palindromic Repeats), así como su proteína asociada Cas (CRISPR associated), especialmente CRISPR-Cas12a y CRISPR-Cas13a, se han caracterizado por su sensibilidad, especificidad, alta resolución de base y programabilidad sobre ácido nucleico, para el reconocimiento de

secuencias, y se han reutilizado para el diagnóstico molecular, abriendo un nuevo camino en la biodetección. Además, son el núcleo de algunas herramientas de diagnóstico sólidas, que están revolucionando la forma en que se pueden detectar los virus. Por lo que CRISPR es reconocido como el sistema inmunitario adaptativo junto con proteínas Cas contra el ADN y virus extraños invasores en bacterias y arqueas.

De acuerdo con la organización de la proteína efectora, los sistemas se dividen en dos clases principales distintas, denominadas Clase 1 y Clase 2, que se subdividen en diferentes tipos y subtipos. Los sistemas de clase 1 incluyen los tipos I, III y IV, que utilizan complejos efectores de múltiples proteínas. A diferencia de la Clase 1, los sistemas de Clase 2 utilizan efectores de una sola proteína. Según las proteínas efectoras, los sistemas de Clase 2 se pueden dividir en tres tipos y varios subtipos, que incluyen sistemas CRISPR de tipo II como Cas9 , sistemas de tipo V como Cas12 (también conocido como cpf1) y sistemas de tipo VI como Cas13.[88]

Los sistemas de repeticiones palindrómicas cortas agrupadas regularmente interespaciadas (CRISPR) son un conjunto de herramientas versátiles de edición de genes que realizan diversas funciones revolucionarias en diversos campos de aplicación, como prácticas agrícolas, industria alimentaria, biotecnología, biomedicina e investigación clínica.

En especial, como método antiviral novedoso de elección, el sistema CRISPR/Cas9 se ha explotado de forma amplia y eficaz para luchar contra los virus infecciosos humanos. Para facilitar la eliminación de virus, el sistema CRISPR/Cas9 ya se ha personalizado para conferir nuevas capacidades antivirales a los animales huéspedes, ya sea modificando el genoma del huésped o dirigiéndose directamente a los factores virales inherentes en forma de ADN. Entre las estrategias, la técnica de edición del genoma de repeticiones palindrómicas cortas agrupadas regularmente interespaciadas (CRISPR)/Cas, como un descubrimiento histórico, ha entrado en el campo de la investigación biomédica y la investigación de la terapia génica, que es muy prometedora para abordar los virus humanos infecciosos graves, el éxito de las modificaciones del genoma a través del aparato CRISPR/Cas9 en células humanas cultivadas ha abierto una nueva ruta para la terapia génica en biomedicina investigación. La edición de genes es un proceso combinado de introducción de escisiones de ADN específicas del sitio por parte de nucleasas y el uso de vías

celulares naturales para reparar las roturas de ADN. Se pueden crear rupturas de doble cadena de ADN exógeno en los genomas por medio de varias plataformas de CRISPR. /Cas sistemas de nucleasas. Luego, las reparaciones de ADN celular iniciadas por lesiones de ADN se completan a través de la vía de reparación dirigida por homología (HDR) con plantillas de reparación o la vía de unión de extremos no homólogos (NHEJ) sin plantillas de reparación.[89]

Por ende, la metodología CRISPR/Cas9 con su diversidad funcional presenta una gran promesa para atacar diversas fases del ciclo de replicación de los virus y tiene la capacidad de llevar a cabo una terapia genética efectiva y sostenida contra los virus que afectan a los seres humanos. En este sentido, en la tabla 8 se resumen, los principales usos de CRISPR para abordar los principales virus infecciosos humanos, como el VIH, el VHB, el VPH y otros como se aprecia en la tabla 8 donde se aplica la tecnología para encontrar vías de señalización en infecciones de estos virus.

Tabla 8. Sitios de orientación de CRISPR-Cas9 en diferentes infecciones virales. Aplicar la tecnología CRISPR-Cas9 para apuntar a los genomas de virus y encontrar vías de señalización que estén involucradas en las infecciones de virus

Virus	Sitio de destino del gen en virus/humano	Modelo	Entrega
Epstein–Barr virus VEB	EBNA1, LMP1, EBNA3C	Líneas celulares de linfoma de Burkitt Célula Raji	Transfección
	BVRF1	Línea celular de cáncer gástrico, SUN719 y YCCEL1	Transfección
	BART5, BART6 o BART16	línea celular de carcinoma gástrico SNU-719	Transducción
Human T-Lymphotropic virus 1 HTLV1	Región pX	ED de células T	Transducción
	RNF8	células HeLa	Electroporación
John cunningham virus	Antígeno T	Línea celular de oligodendroglioma humano, células gliales fetales humanas primarias	Transfección

Lin, H.; Li, G.; Peng, X.; Deng, A.; Ye, L.; Shi, L.; Wang, T.; He, J. The Use of CRISPR/Cas9 as a Tool to Study Human Infectious Viruses. *Front. Cell. Infect. Microbiol.* **2021**, *11* (August), 1–14. https://doi.org/10.3389/fcimb.2021.590989.

Capítulo 8. Herramientas de predicción de estructura viral

Para saber de la base molecular de la función de una proteína es necesario el conocimiento de su estructura nativa, lo que hoy en día es un desafío cada vez mayor, ya que hay una brecha creciente entre las proteínas con estructuras experimentalmente conocidas y proteínas sin estructuras conocidas, frente a la inmensa cantidad de datos, por lo que existe una colección de datos automatizados como herramientas bioinformáticas que ayudan a la determinación de la estructura de una proteína a partir de su secuencia de aminoácidos.[90]

La información 3D es fundamental en la investigación básica, para comprender los mecanismos detrás de la entrada y replicación viral, así como en el diseño de fármacos basados en la estructura, para determinar nuevos objetivos antivirales, o en el desarrollo de vacunas, para estudiar los efectos de nuevas mutaciones en la unión antígeno-anticuerpo, una técnica innovadora, el modelado de proteínas de novo, no requiere una estructura de plantilla y puede complementar los métodos existentes. Los modelos basados en plantillas suelen ser más precisos que los de novo; sin embargo, la primera técnica depende de estructuras previamente resueltas de proteínas homólogas o complejos de proteínas, mientras que la última puede aplicarse a proteínas nuevas. El modelado de formas 3D de las proteínas virales y sus complejos clave brinda conocimiento estructural del virus varios meses críticos antes de lo que pueden hacerlo los experimentos. Además, una nueva generación de herramientas de modelado de proteínas impulsadas por inteligencia artificial (IA), como AlphaFold2 proporciona una mejora aún mayor en los modelos de proteínas para nuevos virus. Aun así, el modelado de novo debe usarse con precaución y respaldado por experimentos al caracterizar proteínas virales porque su repertorio estructural notablemente diverso podría no capturarse durante el entrenamiento de un método de IA.[91]

Las proteínas juegan un papel crucial tanto en la construcción de estructuras biológicas como en la gestión de procesos bioquímicos en organismos vivos. Las proteínas son polímeros lineales no ramificados de residuos de aminoácidos. Para poseer actividad biológica, las proteínas adoptan estructuras tridimensionales únicas, pliegues), lo que se conoce como el estado nativo. La estructura plegada está determinada por la secuencia de aminoácidos de la proteína estructura primaria y la

formación de la conformación nativa plegada estructura terciaria comienza con un plegado rápido en una estructura secundaria que es una conformación espacial local del esqueleto polipeptídico, estabilizada por enlaces de hidrógeno intramoleculares. Los elementos más comunes de la estructura secundaria son hélices α y láminas β. La llamada estructura cuaternaria es el resultado del ensamblaje de las proteínas plegadas o subunidades de proteínas en complejos proteicos de proteína completamente funcional. El conocimiento de la estructura tridimensional de las proteínas es importante para comprender sus funciones. Un conocimiento detallado de la estructura tridimensional es crucial para el diseño de fármacos basados en la estructura de proteínas. Las estructuras de proteínas determinadas experimentalmente se almacenan en bases de datos, la más grande de ellas es el Protein Data Bank (PDB) disponible públicamente

Una de las técnicas de aprendizaje profundo más populares es la solución de extremo a extremo AlphaFold2 basada en redes neuronales de Alphabet-Google DeepMind, esta técnica es capaz de predecir la distribución de la distancia y la torsión de las proteínas, utilizando esquemas de entrenamiento de estructuras PDB determinadas experimentalmente, secuencias primarias de proteínas y el alineamiento de secuencias múltiples (MSA) de proteínas.[92]

8.1. Modelado de proteínas y estructuras virales

Los virus también tienen aplicaciones biomédicas y biotecnológicas como la terapia génica, la administración de fármacos, así que hoy en día es fundamental tener la compresión de cómo funcionan estos virus para poder usarlos de forma beneficiosa, recordemos que la estructura del virus se compone de una protectora compuesta de proteínas, esta estructura es llamada cápside, ya sea de ADN o ARN, estas cápsides son muy estables y resistentes. Estas cápsides son construidas por autoensamblaje. Por ello es importante explicar por qué medios pasa este autoensamblaje de la cápside viral, ya que permitiría el desarrollo de enfoques innovadores para poder afectar el proceso de ensamblaje y por ende prevenir infecciones virales.[93]

8.2. Docking molecular

El propósito del docking molecular es poder proponer un modelo de unión entre dos moléculas, este método sería de mucha ayuda para la química farmacéutica para la elaboración de nuevos fármacos, este método es muy utilizado para poder predecir la interacción de dos moléculas, generando así un modelo de unión, estas uniones entre otras puede ser entre una molécula pequeña y una macromolécula, este método de docking tiene como base la mecánica molecular, dicha mecánica hace uso de un sistema poliatómico que utiliza la física clásica, los parámetros experimentales, como las cargas, los ángulos de torsión y geométricos, se utilizan para reducir la diferencia entre los datos experimentales y las predicciones de la mecánica molecular, pero debido a las deficiencias y limitaciones de los parámetros experimentales, las ecuaciones matemáticas a menudo se pueden parametrizar sobre la base de cálculos teóricos *ab initio* y semiempíricos de la mecánica cuántica.[73]

8.3. Predicción de interacciones proteína-proteína

Muchas actividades virales dependen de las interacciones proteína-proteína, Estas interacciones son contactos físicos entre proteínas impulsados por las propiedades electromagnéticas de sus componentes, se sabe que las interacciones definen funciones importantes como la transcripción y replicación del ADN además de la señalización celular, sin embargo se han hecho métodos *in vivo* e *in vitro* para su detección de estas interacciones, sin embargo estos son bastante caros, por lo que se ha optado en tomar enfoques computacionales para poder tener estas predicciones de proteína-proteína. Estos métodos se basan en representaciones artesanales de las proteínas involucradas en las interacciones, utilizando diferentes tipos de datos como lo son los datos de secuencia y propiedades fisicoquímicas de los aminoácidos, entre otros.[94]

CONCLUSIONES

• La bioinformática es una herramienta que mediante su software nos ayudará a la hora de estudiar a los virus permitiendo una mayor compresión ya la vez desarrollando estrategias para un mayor control a estos y mejorar el control de enfermedades en la salud pública.

• Las bases de datos y sus respectivos programas informáticos nos brindarán un código y lenguaje informático con funciones administrativas para una mejor logística en recuperar los datos requeridos por igual para garantizar la integridad, seguridad y respaldo de los datos.

• La filogenética nos va a ayudar a la hora de analizar y comparar secuencias moleculares, como el ADN o el ARN viral. Al analizar las secuencias genéticas de diferentes cepas o variantes de un virus, se pueden inferir su relación evolutiva y determinar cómo

• Conocer las técnicas tradicionales de identificación de virus nos ayudará a la hora de poder caracterizar los datos según la clasificación de Baltimore.

• Con ayuda de la metagenómica es posible obtener secuencias del genoma de los diferentes microorganismos, virus en este caso, que componen una comunidad, extrayendo y analizando su ADN de forma global.

• Es posible predecir el hospedero de un virus con un análisis de su secuencia de ADN o ARN con las secuencias espaciadoras del CRISP que son es una familia de secuencias de ADN y se derivan de fragmentos de ADN de bacteriófagos que habían infectado.

PERSPECTIVAS

Con el surgimiento de la bioinformática, este trabajo pretende recopilar, almacenar y analizar los datos biológicos, desde los derivados de la secuenciación genómica, proteómica, metabolómica, entre otros, desarrollando algoritmos o modelos matemáticos para extraer el máximo conocimiento de los datos y aplicarlo directamente a la resolución de problemas biológicos o biomédicos.

Ya que entre los problemas más relevantes que se han visto beneficiados del desarrollo de la genómica y de la bioinformática están, entre muchos otros, el estudio de la identificación del patógeno causante de un brote infeccioso o el descubrimiento de nuevos virus.

BIBLIOGRAFÍA

(1) Luscombe, N. M.; Greenbaum, D.; Gerstein, M. *What Is Bioinformatics? An Introduction and Overview*; 2001; Vol. 10. https://doi.org/10.1055/s-0038-1638103.

(2) Pappas, N.; Roux, S.; Hölzer, M.; Lamkiewicz, K.; Mock, F.; Marz, M.; Dutilh, B. E. Virus Bioinformatics. *Encycl. Virol. Vol. 1-5, Fourth Ed.* **2020**, *1–5* (January), 124–132. https://doi.org/10.1016/B978-0-12-814515-9.00034-5.

(3) Yeh, Y. T.; Gulino, K.; Zhang, Y. H.; Sabestien, A.; Chou, T. W.; Zhou, B.; Lin, Z.; Albert, I.; Lu, H.; Swaminathan, V.; Ghedin, E.; Terrones, M. A Rapid and Label-Free Platform for Virus Capture and Identification from Clinical Samples. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (2), 895–901. https://doi.org/10.1073/pnas.1910113117.

(4) LABETOULLE, M. Epidemiology of Viral Infections. *Acta Ophthalmol.* **2009**, *87* (January), 0–0. https://doi.org/10.1111/j.1755-3768.2009.1323.x.

(5) World Health Organization. *Global Genomic Surveillance Strategy 2022-2032*; 2021.

(6) Singh, G.; Nahirniak, S.; Lamarche, Y.; Fan, E. Laboratory Techniques for Diagnosis of Virus Infections. **2020**, No. January.

(7) WHO. Use of Cell Culture in Virology for Developing Countries in the South-East Asia Region. *World Heal. Organ. Reg. Off. South-East Asia* **2017**, 126.

(8) Hematian, A.; Sadeghifard, N.; Mohebi, R.; Taherikalani, M.; Nasrolahi, A.; Amraei, M.; Ghafourian, S. Traditional and Modern Cell Culture in Virus Diagnosis. *Osong Public Heal. Res. Perspect.* **2016**, *7* (2), 77–82. https://doi.org/10.1016/j.phrp.2015.11.011.

(9) Leland, D. S.; Ginocchio, C. C. Role of Cell Culture for Virus Detection in the Age of Technology. *Clin. Microbiol. Rev.* **2007**, *20* (1), 49–78. https://doi.org/10.1128/CMR.00002-06.

(10) He, B.; Chen, G.; Zeng, Y. Three-Dimensional Cell Culture Models for Investigating Human Viruses. *Virol. Sin.* **2016**, *31* (5), 363–379. https://doi.org/10.1007/s12250-016-3889-z.

(11) Dolskiy, A. A.; Grishchenko, I. V.; Yudkin, D. V. Cell Cultures for Virology: Usability, Advantages, and Prospects. *Int. J. Mol. Sci.* **2020**, *21* (21), 1–23. https://doi.org/10.3390/ijms21217978.

(12) Fong, C. K. Y. Electron Microscopy for the Rapid Detection and Identification of Viruses from Clinical Specimens. *Yale J. Biol. Med.* **1989**, *62* (2), 115–130.

(13) Bagnell, C. R.; Ph, D.; Coons, A. H. Chapter 12 Fluorescence Microscopy. **2000**, 1–8.

(14) Witte, R.; Andriasyan, V.; Georgi, F.; Yakimovich, A.; Greber, U. F. Concepts in Light Microscopy of Viruses. *Viruses* **2018**, *10* (4), 1–31. https://doi.org/10.3390/v10040202.

(15) Diagnostic Techniques: Serological and Molecular Approaches. **2020**, No. January.

(16) Pathogenic Viruses: Molecular Detection and Characterization. **2020**, No. January.

(17) Watzinger, F.; Ebner, K.; Lion, T. Detection and Monitoring of Virus Infections by Real-Time PCR. *Mol. Aspects Med.* **2006**, *27* (2–3), 254–298. https://doi.org/10.1016/j.mam.2005.12.001.

(18) Ongay-Larios, L.; Códiz Huerta, G. Secuenciación de ADN Por El Método de Terminación de La Cadena de Sanger. *Mensaje Bioquímico* **2021**, *45*, 23–34.

(19) Kumar, A.; Murthy, S.; Kapoor, A. Evolution of Selective-Sequencing Approaches for Virus Discovery and Virome Analysis. **2020**, No. January.

(20) Ho, T.; Tzanetakis, I. E. Development of a Virus Detection and Discovery Pipeline Using next Generation Sequencing. **2014**, *473*, 54–60.

(21) Koonin, E. V. The Baltimore Classi Fi Cation of Viruses 50 Years Later : How Does It Stand in the Light of Virus Evolution ? **2021**, *85* (3), 1–19.

(22) Gelderblom, H. R. Structure and Classification of Viruses. *Med. Microbiol.* **1996**, No. January 1996.

(23) Virus Databases. **2020**, No. January.

(24) ZHDANOV, V. M.; GAIDAMOVICH, S. I. On the Classification and Nomenclature of Viruses. *Vopr. Virusol.* **1962**, *7*, 749–754.

(25) Lefkowitz, E. J.; Dempsey, D. M.; Hendrickson, R. C.; Orton, R. J.; Siddell, S. G.; Smith, D. B. Virus Taxonomy: The Database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **2018**, *46* (D1), D708–D717. https://doi.org/10.1093/nar/gkx932.

(26) Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A Knowledge Resource to Understand Virus Diversity. *Nucleic Acids Res.* **2011**, *39* (SUPPL. 1), 576–582.

https://doi.org/10.1093/nar/gkq901.

(27) Pickett, B. E.; Greer, D. S.; Zhang, Y.; Stewart, L.; Zhou, L.; Sun, G.; Gu, Z.; Kumar, S.; Zaremba, S.; Larsen, C. N.; Jen, W.; Klem, E. B.; Scheuermann, R. H. Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community. *Viruses* **2012**, *4* (11), 3209–3226. https://doi.org/10.3390/v4113209.

(28) Mihara, T.; Nishimura, Y.; Shimizu, Y.; Nishiyama, H.; Yoshikawa, G.; Uehara, H.; Hingamp, P.; Goto, S.; Ogata, H. Linking Virus Genomes with Host Taxonomy. *Viruses* **2016**, *8* (3), 10–15. https://doi.org/10.3390/v8030066.

(29) Lamy-besnier, Q.; Brancotte, B.; Debarbieux, L.; Computationnelle, B.; Pasteur, I.; F-, P. Viral Host Range Database , an Online Tool for Recording , Analyzing and Disseminating Virus – Host Interactions. **2021**, *37* (February), 2798–2801. https://doi.org/10.1093/bioinformatics/btab070.

(30) Brister, J. R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Res.* **2015**, *43* (D1), D571–D577. https://doi.org/10.1093/nar/gku1207.

(31) Paez-Espino, D.; Chen, I. M. A.; Palaniappan, K.; Ratner, A.; Chu, K.; Szeto, E.; Pillay, M.; Huang, J.; Markowitz, V. M.; Nielsen, T.; Huntemann, M.; Reddy, T. B. K.; Pavlopoulos, G. A.; Sullivan, M. B.; Campbell, B. J.; Chen, F.; McMahon, K.; Hallam, S. J.; Denef, V.; Cavicchioli, R.; Caffrey, S. M.; Streit, W. R.; Webster, J.; Handley, K. M.; Salekdeh, G. H.; Tsesmetzis, N.; Setubal, J. C.; Pope, P. B.; Liu, W. T.; Rivers, A. R.; Ivanova, N. N.; Kyrpides, N. C. IMG/VR: A Database of Cultured and Uncultured DNA Viruses and Retroviruses. *Nucleic Acids Res.* **2017**, *45* (D1), D457–D465. https://doi.org/10.1093/nar/gkw1030.

(32) Camargo, A. P.; Nayfach, S.; Chen, I. M. A.; Palaniappan, K.; Ratner, A.; Chu, K.; Ritter, S. J.; Reddy, T. B. K.; Mukherjee, S.; Schulz, F.; Call, L.; Neches, R. Y.; Woyke, T.; Ivanova, N. N.; Eloe-Fadrosh, E. A.; Kyrpides, N. C.; Roux, S. IMG/VR v4: An Expanded Database of Uncultivated Virus Genomes within a Framework of Extensive Functional, Taxonomic, and Ecological Metadata. *Nucleic Acids Res.* **2023**, *51* (D1), D733–D743. https://doi.org/10.1093/nar/gkac1037.

(33) Wu, H.; Fu, P.; Fu, Q.; Zhang, Z.; Zheng, H.; Mao, L.; Li, X.; Yu, F.; Peng, Y. Plant Virus Database: A Resource for Exploring the Diversity of Plant Viruses

and Their Interactions with Hosts. *bioRxiv* **2022**, 2022.03.20.485054.

(34) Roux, S.; Adriaenssens, E. M.; Dutilh, B. E.; Koonin, E. V.; Kropinski, A. M.; Krupovic, M.; Kuhn, J. H.; Lavigne, R.; Brister, J. R.; Varsani, A.; Amid, C.; Aziz, R. K.; Bordenstein, S. R.; Bork, P.; Breitbart, M.; Cochrane, G. R.; Daly, R. A.; Desnues, C.; Duhaime, M. B.; Emerson, J. B.; Enault, F.; Fuhrman, J. A.; Hingamp, P.; Hugenholtz, P.; Hurwitz, B. L.; Ivanova, N. N.; Labonté, J. M.; Lee, K. B.; Malmstrom, R. R.; Martinez-Garcia, M.; Mizrachi, I. K.; Ogata, H.; Páez-Espino, D.; Petit, M. A.; Putonti, C.; Rattei, T.; Reyes, A.; Rodriguez-Valera, F.; Rosario, K.; Schriml, L.; Schulz, F.; Steward, G. F.; Sullivan, M. B.; Sunagawa, S.; Suttle, C. A.; Temperton, B.; Tringe, S. G.; Thurber, R. V.; Webster, N. S.; Whiteson, K. L.; Wilhelm, S. W.; Wommack, K. E.; Woyke, T.; Wrighton, K. C.; Yilmaz, P.; Yoshida, T.; Young, M. J.; Yutin, N.; Allen, L. Z.; Kyrpides, N. C.; Eloe-Fadrosh, E. A. Minimum Information about an Uncultivated Virus Genome (MIUVIG). *Nat. Biotechnol.* **2019**, *37* (1), 29–37. https://doi.org/10.1038/nbt.4306.

(35) Wu, L.; Pappas, N.; Wijesekara, Y.; Piedade, G. J.; Corina, P. D.; Dutilh, B. E. Benchmarking Bioinformatic Virus Identification Tools Using Real-World Metagenomic Data across Biomes. **2023**.

(36) Bassi, C.; Guerriero, P.; Pierantoni, M.; Callegari, E.; Sabbioni, S. Novel Virus Identification through Metagenomics: A Systematic Review. *Life* **2022**, *12* (12). https://doi.org/10.3390/life12122048.

(37) Rose, R.; Constantinides, B.; Tapinos, A.; Robertson, D. L.; Prosperi, M. Challenges in the Analysis of Viral Metagenomes. *Virus Evol.* **2016**, *2* (2), 1–11. https://doi.org/10.1093/ve/vew022.

(38) Andrade-Martínez, J. S.; Camelo Valera, L. C.; Chica Cárdenas, L. A.; Forero-Junco, L.; López-Leal, G.; Moreno-Gallego, J. L.; Rangel-Pineros, G.; Reyes, A. Computational Tools for the Analysis of Uncultivated Phage Genomes. *Microbiol. Mol. Biol. Rev.* **2022**, *86* (2). https://doi.org/10.1128/mmbr.00004-21.

(39) Orton, R. J.; Gu, Q.; Hughes, J.; Maabar, M.; Modha, S.; Vattipally, S. B.; Wilkie, G. S.; Davison, A. J. Bioinformatics Tools for Analysing Viral Genomic Data. *OIE Rev. Sci. Tech.* **2016**, *35* (1), 271–285. https://doi.org/10.20506/rst.35.1.2432.

(40) Roux, S.; Enault, F.; Hurwitz, B. L.; Sullivan, M. B. VirSorter : Mining Viral Signal from Microbial Genomic Data. **2015**, 1–20. https://doi.org/10.7717/peerj.985.

(41) Wan, Y.; Renner, D. W.; Albert, I.; Szpara, M. L. VirAmp: A Galaxy-Based Viral

Genome Assembly Pipeline. *Gigascience* **2015**, *4* (1). https://doi.org/10.1186/s13742-015-0060-y.

(42) Wommack, K. E.; Bhavsar, J.; Polson, S. W.; Chen, J.; Dumas, M.; Srinivasiah, S.; Furman, M.; Jamindar, S.; Nasko, D. J. VIROME : A Standard Operating Procedure for Analysis of Viral Metagenome Sequences. **2012**, 427–439. https://doi.org/10.4056/sigs.2945050.

(43) Arndt, D.; Grant, J. R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D. S. PHASTER : A Better , Faster Version of the PHAST Phage Search Tool. **2016**, *44* (May), 16–21. https://doi.org/10.1093/nar/gkw387.

(44) Guo, J.; Bolduc, B.; Zayed, A. A.; Varsani, A.; Dominguez-Huerta, G.; Delmont, T. O.; Pratama, A. A.; Gazitúa, M. C.; Vik, D.; Sullivan, M. B.; Roux, S. VirSorter2: A Multi-Classifier, Expert-Guided Approach to Detect Diverse DNA and RNA Viruses. *Microbiome* **2021**, *9* (1), 1–13. https://doi.org/10.1186/s40168-020-00990-y.

(45) Ponsero, A. J.; Hurwitz, B. L. The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes. **2019**, *10* (April), 1–6. https://doi.org/10.3389/fmicb.2019.00806.

(46) Ren, J.; Ahlgren, N. A.; Lu, Y. Y.; Fuhrman, J. A.; Sun, F. VirFinder: A Novel k-Mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data. *Microbiome* **2017**, *5* (1), 69. https://doi.org/10.1186/s40168-017-0283-5.

(47) Ren, J.; Song, K.; Deng, C.; Ahlgren, N. A.; Fuhrman, J. A.; Li, Y.; Xie, X.; Poplin, R.; Sun, F. Identifying Viruses from Metagenomic Data Using Deep Learning. *Quant. Biol.* **2020**, *8* (1), 64–77. https://doi.org/10.1007/s40484-019-0187-4.

(48) Zhao, G.; Krishnamurthy, S.; Cai, Z.; Popov, V. L.; Travassos da Rosa, A. P.; Guzman, H.; Cao, S.; Virgin, H. W.; Tesh, R. B.; Wang, D. Identification of Novel Viruses Using VirusHunter -- an Automated Data Analysis Pipeline. *PLoS One* **2013**, *8* (10), 1–11. https://doi.org/10.1371/journal.pone.0078470.

(49) Naccache, S. N.; Angeles, H. L.; Veeraraghavan, N.; Children, R.; Samayoa, E.; Francisco, S. A Cloud-Compatible Bioinformatics Pipeline for Ultrarapid Pathogen Identification from next-Generation Sequencing of Clinical Samples A Cloud-Compatible Bioinformatics Pipeline for Ultrarapid Pathogen Identification from next-Generation Sequencing of Clin. **2014**, No. September. https://doi.org/10.1101/gr.171934.113.

(50) Li, Y.; Wang, H.; Nie, K.; Zhang, C.; Zhang, Y.; Wang, J.; Niu, P. OPEN VIP : An

Integrated Pipeline for Metagenomics of Virus Identification and Discovery. **2016**, No. September 2015, 1–10. https://doi.org/10.1038/srep23774.

(51) Zhao, G.; Wu, G.; Lim, E. S.; Droit, L.; Krishnamurthy, S.; Barouch, D. H.; Virgin, H. W.; Wang, D. VirusSeeker, a Computational Pipeline for Virus Discovery and Virome Composition Analysis. *Virology* **2017**, *503* (January), 21–30. https://doi.org/10.1016/j.virol.2017.01.005.

(52) Rampelli, S.; Soverini, M.; Turroni, S.; Quercia, S.; Biagi, E.; Brigidi, P.; Candela, M. ViromeScan: A New Tool for Metagenomic Viral Community Profiling. *BMC Genomics* **2016**, *17* (1), 1–10. https://doi.org/10.1186/s12864-016-2446-3.

(53) Yamashita, A.; Sekizuka, T.; Kuroda, M. VirusTAP: Viral Genome-Targeted Assembly Pipeline. *Front. Microbiol.* **2016**, *7* (FEB), 1–5. https://doi.org/10.3389/fmicb.2016.00032.

(54) Rampelli, S.; Soverini, M.; Turroni, S.; Quercia, S.; Biagi, E.; Brigidi, P.; Candela, M. ViromeScan : A New Tool for Metagenomic Viral Community Profiling. **2016**, 1–9. https://doi.org/10.1186/s12864-016-2446-3.

(55) Thorburn, F.; Maabar, M.; Davison, A. J.; Vu, M.; Murcia, P. R.; Gunson, R.; Palmarini, M.; Hughes, J. DisCVR : Rapid Viral Diagnosis from High-Throughput Sequencing Data. **2019**, *5* (2), 1–8. https://doi.org/10.1093/ve/vez033.

(56) Lefebvre, M. The VirAnnot Pipeline : A Resource for Automated Viral Diversity Estimation and Operational Taxonomy Units Assignation for Virome Sequencing Data. **2019**, No. August, 256–259.

(57) Laffy, P. W.; Wood-Charlson, E. M.; Turaev, D.; Weynberg, K. D.; Botté, E. S.; Van Oppen, M. J. H.; Webster, N. S.; Rattei, T. HoloVir: A Workflow for Investigating the Diversity and Function of Viruses in Invertebrate Holobionts. *Front. Microbiol.* **2016**, *7* (JUN), 1–15. https://doi.org/10.3389/fmicb.2016.00822.

(58) Tithi, S. S.; Aylward, F. O.; Jensen, R. V.; Zhang, L. FastViromeExplorer: A Pipeline for Virus and Phage Identification and Abundance Profiling in Metagenomics Data. *PeerJ* **2018**, *2018* (1), 1–18. https://doi.org/10.7717/peerj.4227.

(59) Baizan-Edge, A.; Cock, P.; MacFarlane, S.; McGavin, W.; Torrance, L.; Jones, S. Kodoja: A Workflow for Virus Detection in Plants Using k-Mer Analysis of RNA-Sequencing Data. *J. Gen. Virol.* **2019**, *100* (3), 533–542. https://doi.org/10.1099/jgv.0.001210.

(60) Zheng, Y.; Gao, S.; Padmanabhan, C.; Li, R.; Galvez, M.; Gutierrez, D.; Fuentes, S.; Ling, K.; Kreuze, J.; Fei, Z. VirusDetect:Anautomatedpipelineforefficientvirusdiscoveryusingdeep SequencingofsmallRNAs. **2017**, *500* (August 2016), 130–138.

(61) Plyusnin, I.; Kant, R.; Jaä¨skelaïnen, A. J.; Sironen, T.; Holm, L.; Vapalahti, O.; Smura, T. Novel NGS Pipeline for Virus Discovery from a Wide Spectrum of Hosts and Sample Types. *Virus Evol.* **2020**, *6* (2), 1–10. https://doi.org/10.1093/ve/veaa091.

(62) Smits, S. L.; Bodewes, R.; Ruiz-Gonzalez, A.; Baumgärtner, W.; Koopmans, M. P.; Osterhaus, A. D. M. E.; Schürch, A. C. Assembly of Viral Genomes from Metagenomes. *Front. Microbiol.* **2014**, *5* (DEC), 1–10. https://doi.org/10.3389/fmicb.2014.00714.

(63) Andrade-martínez, J. S.; Camelo, C.; Chica, A.; Forero-junco, L.; López-leal, G.; Moreno-gallego, J. L.; Rangel-pineros, G. Computational Tools for the Analysis of Uncultivated Phage Genomes.

(64) Yoshida, R.; Page, R. Phylogenetic Analysis and Molecular Evolution (PAME). *Primus* **2022**, *32* (3), 386–415. https://doi.org/10.1080/10511970.2021.1919257.

(65) Nasir, A.; Caetano-Anollés, G. A Phylogenomic Data-Driven Exploration of Viral Origins and Evolution. *Sci. Adv.* **2015**, *1* (8). https://doi.org/10.1126/sciadv.1500527.

(66) Catanach, T. A.; Sweet, A. D.; Nguyen, N. P. D.; Peery, R. M.; Debevec, A. H.; Thomer, A. K.; Owings, A. C.; Boyd, B. M.; Katz, A. D.; Soto-Adames, F. N.; Allen, J. M. Fully Automated Sequence Alignment Methods Are Comparable to, and Much Faster than, Traditional Methods in Large Data Sets: An Example with Hepatitis B Virus. *PeerJ* **2019**, *2019* (1). https://doi.org/10.7717/peerj.6142.

(67) Chowdhury, B.; Garai, G. Genomics A Review on Multiple Sequence Alignment from the Perspective of Genetic Algorithm. **2017**, *109*, 419–431.

(68) Chao, J.; Tang, F. Developments in Algorithms for Sequence Alignment : A Review. **2022**, 1–13.

(69) Ponce-de-Leon-Senti, E.; Diaz, E.; Guardado-Muro, H.; Cuellar-Garrido, D.; Martinez-Guerra, J. J.; Torres-Soto, A.; Torres-Soto, D.; Hernandez-Aguirre, A. A Distance Measure for Building Phylogenetic Trees: A First Approach. *Res. Comput. Sci.* **2017**, *139* (1), 149–162. https://doi.org/10.13053/rcs-139-1-12.

(70) Duchen, P. Bayesian Inference. **2021**, *4*, 81–89.

(71) Pérez-Losada, M.; Arenas, M.; Galán, J. C.; Palero, F.; González-Candelas, F. Recombination in Viruses: Mechanisms, Methods of Study, and Evolutionary Consequences. *Infect. Genet. Evol.* **2015**, *30*, 296–307. https://doi.org/10.1016/j.meegid.2014.12.022.

(72) Martin, D. P.; Murrell, B.; Golden, M.; Khoosal, A.; Muhire, B. RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes. *Virus Evol.* **2015**, *1* (1), 1–5. https://doi.org/10.1093/ve/vev003.

(73) Molecular Docking Current Advances and Challenges.Pdf.

(74) Perfectti, F. La Huella Genética de La Selección Natural. **2014**, No. May.

(75) *The Phylogenetic Handbook*.

(76) Pond, S. L. K.; Frost, S. D. W. Not So Different After All : A Comparison of Methods for Detecting Amino Acid Sites Under Selection. **2004**. https://doi.org/10.1093/molbev/msi105.

(77) Perdoncini Carvalho, C.; Ren, R.; Han, J.; Qu, F. Natural Selection, Intracellular Bottlenecks of Virus Populations, and Viral Superinfection Exclusion. *Annu. Rev. Virol.* **2022**, *9*, 121–137. https://doi.org/10.1146/annurev-virology-100520-114758.

(78) Weaver, S.; Shank, S. D.; Spielman, S. J.; Li, M.; Muse, S. V.; Kosakovsky Pond, S. L. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol. Biol. Evol.* **2018**, *35* (3), 773–777. https://doi.org/10.1093/molbev/msx335.

(79) Douam, F.; Fusil, F.; Enguehard, M.; Dib, L.; Nadalin, F.; Schwaller, L.; Hrebikova, G.; Mancip, J.; Mailly, L.; Montserret, R.; Ding, Q.; Maisse, C.; Carlot, E.; Xu, K.; Verhoeyen, E.; Baumert, T. F.; Ploss, A.; Carbone, A.; Cosset, F. L.; Lavillette, D. *A Protein Coevolution Method Uncovers Critical Features of the Hepatitis C Virus Fusion Mechanism*; 2018; Vol. 14. https://doi.org/10.1371/journal.ppat.1006908.

(80) Mock, F.; Viehweger, A.; Barth, E.; Marz, M. VIDHOP, Viral Host Prediction with Deep Learning. *Bioinformatics* **2021**, *37* (3), 318–325. https://doi.org/10.1093/bioinformatics/btaa705.

(81) Villarroel, J.; Kleinheinz, K. A.; Jurtz, V. I.; Zschach, H.; Lund, O.; Nielsen, M.; Larsen, M. V. HostPhinder: A Phage Host Prediction Tool. *Viruses* **2016**, *8* (5), 1–22. https://doi.org/10.3390/v8050116.

(82) Liu, D.; Young, F.; Robertson, D. L.; Yuan, K. Prediction of Virus-Host Association Using Protein Language Models and Multiple Instance Learning Author Summary. **2023**, 1–23.

(83) Poon, A. F. Y.; Lewis, F. I.; Frost, S. D. W.; Kosakovsky Pond, S. L. Spidermonkey: Rapid Detection of Co-Evolving Sites Using Bayesian Graphical Models. *Bioinformatics* **2008**, *24* (17), 1949–1950. https://doi.org/10.1093/bioinformatics/btn313.

(84) Pons, J. C.; Paez-Espino, D.; Riera, G.; Ivanova, N.; Kyrpides, N. C.; Llabrés, M. VPF-Class: Taxonomic Assignment and Host Prediction of Uncultivated Viruses Based on Viral Protein Families. *Bioinformatics* **2021**, *37* (13), 1805–1813. https://doi.org/10.1093/bioinformatics/btab026.

(85) Zielezinski, A.; Deorowicz, S.; Gudyś, A. PHIST: Fast and Accurate Prediction of Prokaryotic Hosts from Metagenomic Viral Sequences. *Bioinformatics* **2022**, *38* (5), 1447–1449. https://doi.org/10.1093/bioinformatics/btab837.

(86) Coutinho, F. H.; Zaragoza-Solas, A.; López-Pérez, M.; Barylski, J.; Zielezinski, A.; Dutilh, B. E.; Edwards, R.; Rodriguez-Valera, F. RaFAH: Host Prediction for Viruses of Bacteria and Archaea Based on Protein Content. *Patterns* **2021**, *2* (7). https://doi.org/10.1016/j.patter.2021.100274.

(87) Asim, M. N.; Fazeel, A.; Ibrahim, M. A.; Dengel, A.; Ahmed, S. MP-VHPPI: Meta Predictor for Viral Host Protein-Protein Interaction Prediction in Multiple Hosts and Viruses. *Front. Med.* **2022**, *9*. https://doi.org/10.3389/fmed.2022.1025887.

(88) Roux, S.; Páez-Espino, D.; Chen, I. M. A.; Palaniappan, K.; Ratner, A.; Chu, K.; Reddy, T.; Nayfach, S.; Schulz, F.; Call, L.; Neches, R. Y.; Woyke, T.; Ivanova, N. N.; Eloe-Fadrosh, E. A.; Kyrpides, N. C. IMG/VR v3: An Integrated Ecological and Evolutionary Framework for Interrogating Genomes of Uncultivated Viruses. *Nucleic Acids Res.* **2021**, *49* (D1), D764–D775. https://doi.org/10.1093/nar/gkaa946.

(89) Lin, H.; Li, G.; Peng, X.; Deng, A.; Ye, L.; Shi, L.; Wang, T.; He, J. The Use of CRISPR/Cas9 as a Tool to Study Human Infectious Viruses. *Front. Cell. Infect. Microbiol.* **2021**, *11* (August), 1–14. https://doi.org/10.3389/fcimb.2021.590989.

(90) Pavlopoulou, A.; Michalopoulos, I. State-of-the-Art Bioinformatics Protein Structure Prediction Tools (Review). *Int. J. Mol. Med.* **2011**, *28* (3), 295–310. https://doi.org/10.3892/ijmm.2011.705.

(91) Narykov, O.; Srinivasan, S.; Korkin, D. Computational Protein Modeling and the

next Viral Pandemic. *Nat. Methods* **2021**, *18* (5), 444–445. https://doi.org/10.1038/s41592-021-01144-0.

(92) Gutnik, D.; Evseev, P.; Miroshnikov, K.; Shneider, M. Using AlphaFold Predictions in Viral Research. *Curr. Issues Mol. Biol.* **2023**, *45* (4), 3705–3732. https://doi.org/10.3390/cimb45040240.

(93) May, E. R.; Arora, K.; Mannige, R. V; Nguyen, H. D.; Iii, C. L. B. Multiscale Modeling of Virus Structure , Assembly , and Dynamics Multiscale Modeling of Virus Structure , Assembly , and Dynamics. **2012**, No. July 2017. https://doi.org/10.1007/978-1-4614-2146-7.

(94) Gonzalez-Lopez, F.; Morales-Cordovilla, J. A.; Villegas-Morcillo, A.; Gomez, A. M.; Sanchez, V. End-to-End Prediction of Protein-Protein Interaction Based on Embedding and Recurrent Neural Networks. *Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018* **2019**, 2344–2350. https://doi.org/10.1109/BIBM.2018.8621328.